POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Evaluation of Consciousness in patients through EEG signals: a Pattern Recognition and Feature Engineering approach

Supervisor
Luca Mesin

Co-Supervisor
Hossein Ahmadi

Candidate
Francesco Serracca

A.Y. 2024/2025

Contents

List of Figures

List of Tables

Chapter 1

Abstract

This thesis deals with the analysis of states of consciousness in patients through EEG (electroencephalography) signals, using pattern recognition and feature engineering . This work aims to implement classification and assessment techniques of DOC patients through brain signals. EEG, due to its high temporal resolution and non-invasive nature, has been chosen as the main technique for brain data acquisition. The process involves several steps, including channel selection, preliminary signal pre-processing, artefact detection, feature extraction in both time and frequency domains, and the use of clustering and classification algorithms for data analysis. Among the classification methods used are Support Vector Machine, Random Forest and logistic regression. The results show the feasibility of these approaches in distinguishing between different states of consciousness, paving the way for future research on advanced methodologies to monitor and assess consciousness in clinical settings.

Chapter 2

Introduction

2.1 Nervous system

For Nervous system we mean, from an anatomical and functional point of view, the unit characterized by highly specialized tissue responsible for processing bio-electrical signals. From an anatomical point of view, it is divided in Central Nervous System (CNS) and Peripheral Nervous System (PNS).

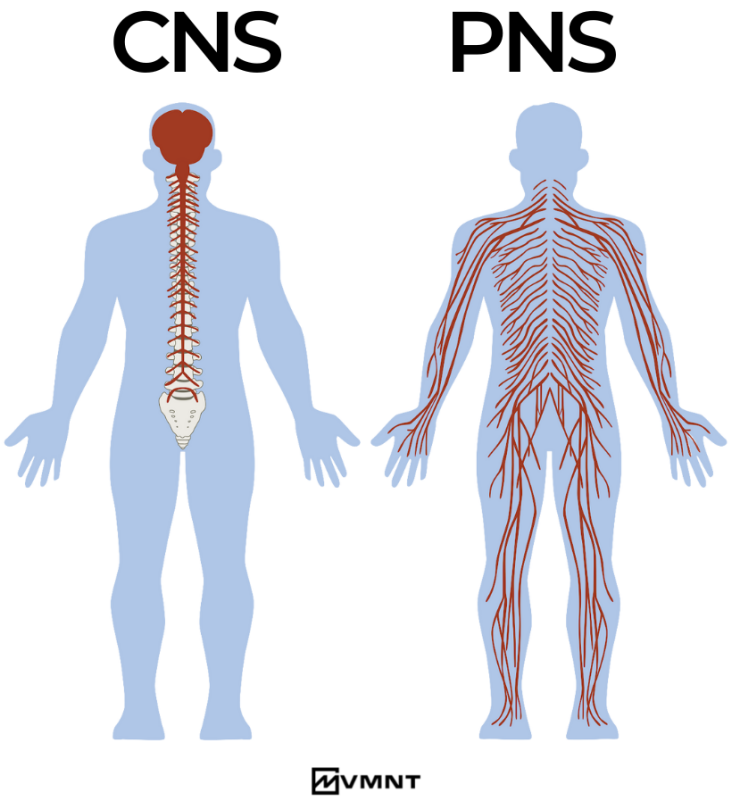

Figure 2.1: Central Nervous System (CNS) and Peripheral Nervous System (PNS) composition.

- **Central Nervous System**: it is enclosed in the skull as regards the encephalon, and in the vertebral canal if we consider the spinal cord. It has the task to collect and integrate the received information from the external environment by the spinal nerves, cranial nerves but also by the senses; furthermore, it coordinates all parts of the body by sending them responses to stimuli.

- **Peripheral Nervous System**: it is made up of peripheral nervous structures such as ganglia, nerves, sensory receptors (thermoreceptors, proprioceptors, mechanoreceptors, odor receptors, taste receptors) and specialized organs such as the eye and the auditory apparatus.

In this chapter, we focus on the central nervous system, the organs that compose it, their composition and the functions they perform.

The contents of the following paragraphs regarding notions of anatomy are taken, unless otherwise specified through further bibliographical references, from the book "Principles of Neural Science" by Eric R. Kandel [9] and from the consultation of the site `https://it.wikipedia.org/`.

2.2 Brain

The brain or encephalon is the principal organ of the CNS and it is involved in a wide range of processes, from learning to memory up to sensory functions such as vision and hearing. It is one of the bigger and more complex organs of our body and it constitutes the hub of the human neural system. It is housed and protected totally within the skull and inside contains numerous structures and billions of cells. As we can also see from the Figure 2.2 it is composed of three organs :

- Cerebrum
- Brainstem
- Cerebellum

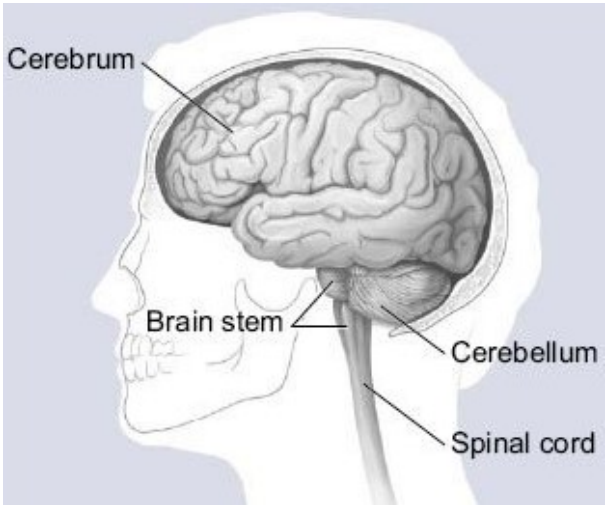

Figure 2.2: Principals organs that compose the Brain.

2.3 Cerebrum

The human cerebrum is the principal organ of the CNS and it is present in all vertebrates organisms and most invertebrates. Its weigh is about 1400 grams, has a volume between 1100 and 1300 cm^3, and contains 10^{12} neurons. From an embryological point of view it is the result of the development of the prosencephalon and derives from the diencephalic and telencephalic vesicles. Anatomically, in fact, it is made up of the diencephalon and telencephalon. The diencephalon is smaller and is wrapped superiorly and laterally by the telencephalon. It aggregates the thalamus, epithalamus, metathalamus, hypothalamus, and subthalamus. It performs key roles such as regulation of body temperature, appetite control, management of cardiac rhythms, sensory perception and transfer of sensory information. It is divided in two hemispheres, the two halves of the brain, divided by the median longitudinal sulcus (interhemispheric fissure).
In turn, each hemisphere is divided into lobes, the main ones are four:

- Frontal Lobe
- Temporal Lobe
- Parietal Lobe
- Occipital Lobe

2.4 Brainstem

The brainstem is situated at the base of the brain and it is made up of of the midbrain, pons, and medulla oblongata (arranged as can be seen in Figure 2.3). The brainstem contains nerve fibers that descend to and ascend from the spinal cord; it also houses several motor and sensory nuclei, which are groups of neurons that further process these signals. The most numerous of these are located within the pons and are collectively known as "pontine nuclei". The basal ganglia are closely connected with the cerebral cortex and play a critical role in movement. Both Parkinson's disease and Huntigton's chorea are due to pathology in the basal ganglia. Ascending towards the skull, the spinal cord progressively changes and becomes more complex, finally articulating as the medulla oblongata or bulb. It contains the centers for the regulation of visceral, respiratory and blood pressure functions.
The **medulla oblongata** is an elongated structure located in the lower part of the brainstem. Information relating to taste, touch and mainly hearing pass from the bulb.
Pons is a large protuberance of the brainstem, which forms part of the floor of the fourth cerebral ventricle, and joins the cerebellum posteriorly via the middle cerebellar peduncle.
Midbrain is the tallest structure of the brainstem. It is involved in various brain functions, including vision (through the regulation of some extrinsic and intrinsic muscles of the eyes), the planning and regulation of muscle movements, as well as reward mechanisms through the substantia nigra, associated with the motor pathways of the basal nuclei. Dopamine, the neurotransmitter also known as "neurohormone" is produced in the substantia nigra and plays a role in the development of motivation, sense of satisfaction and gratification.

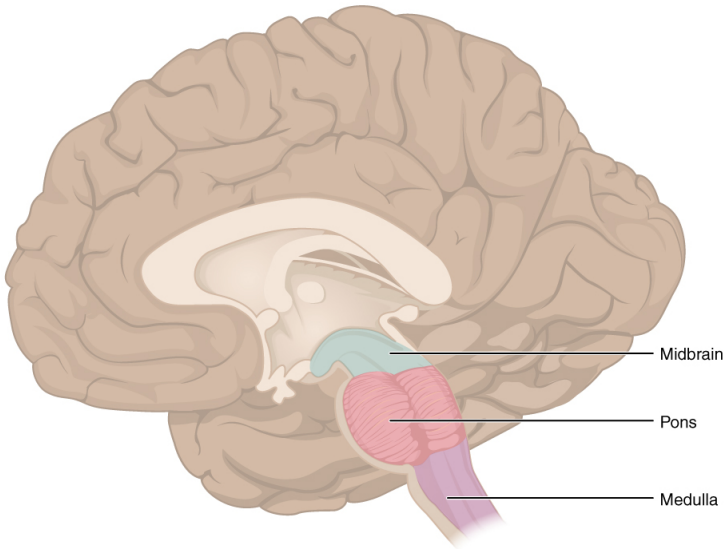

Figure 2.3: Sagittal section of the brain showing the three sections of the brainstem.

2.5 Cerebellum

The cerebellum, whose name comes from the Latin term for "little brain" is situated beneath the posterior part of the cerebral hemispheres (Figure 2.4). It plays a key role in generating smooth, coordinated movements and is also involved in motor learning and adaptation. Although the cerebellum does not have direct connections to the spinal cord, it indirectly influences movement through its connections to the cerebrum and brainstem. People with disorders of the cerebellum are still able to move, but their movements lack normal coordination; these characteristic deficits are known collectively as ataxia.

2.6 Diencephalon

The diencephalon is the central part of the brain situated between the telencephalon and the midbrain (see Figure 2.4). It plays a crucial role in many fundamental functions of the central nervous system, maintaining homeostasis and responding to changes in the external environment. The diencephalon consists of several structures, including the thalamus, hypothalamus, epithalamus, and subthalamus.

The **thalamus** is a large nucleus that acts as a relay station for most sensory information heading to the cerebral cortex. It receives input from nearly all sensory systems (except olfaction) and sends these signals to the appropriate cortical areas for processing.
Additionally, the thalamus is involved in the regulation of sleep, wakefulness, and consciousness.

The **hypothalamus** is a small but vital region of the diencephalon that controls many essential bodily functions. It regulates the endocrine system by controlling the pituitary gland, manages stress responses, and controls body temperature, hunger, thirst, and circadian rhythms, which are the daily cycles of biological activity. The hypothalamus also plays a role in emotions and behavior.

The **epithalamus** includes the pineal gland, which secretes melatonin, a hormone that regulates circadian rhythms and the sleep-wake cycle. The epithalamus also helps regulate certain autonomic functions.

The **subthalamus**, located beneath the thalamus, contains structures such as the subthalamic nucleus. It is involved in motor control and has close connections with the basal ganglia, which are essential for the coordination and modulation of movement. The spinal cord, which extends from the brain along the spine, serves as a "highway" for communication between the brain and the rest of the body.

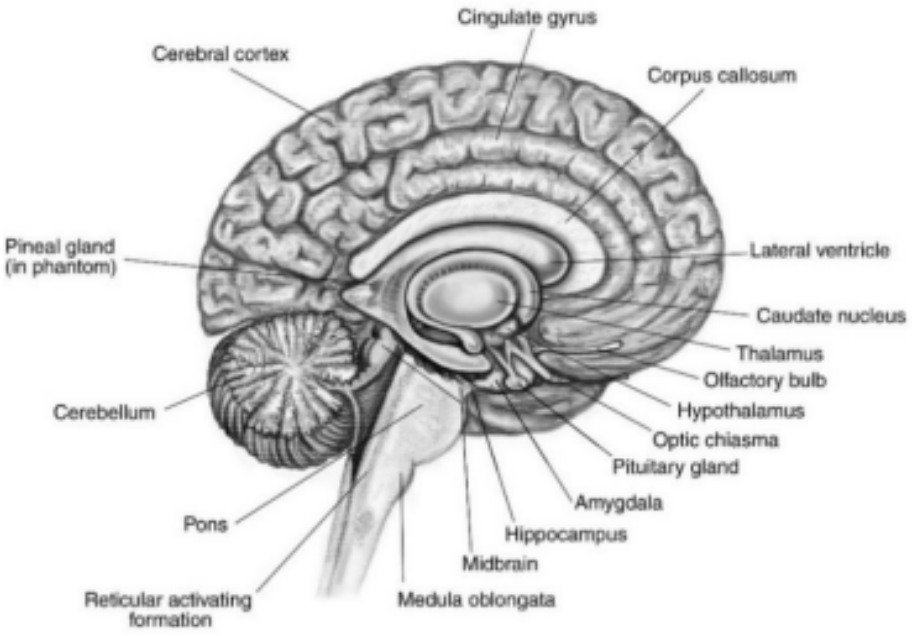

Figure 2.4: Median sagittal view of the brain.

2.7 Meninges

The meninges are a system of membranes located inside the skull that cover the central nervous system and protect the brain and spinal cord. Canonically there are three membranous connective layers that appear as concentric sheets Figure 1.5, but recent studies have led to considering a fourth. Respectively, from the outside to the inside, they are called:

- Dura mater (or dura meninges or pachymeninges)
- Arachnoid
- Lymphatic-like subarachnoid (SLYM)
- Pia mater (or pia meninge)

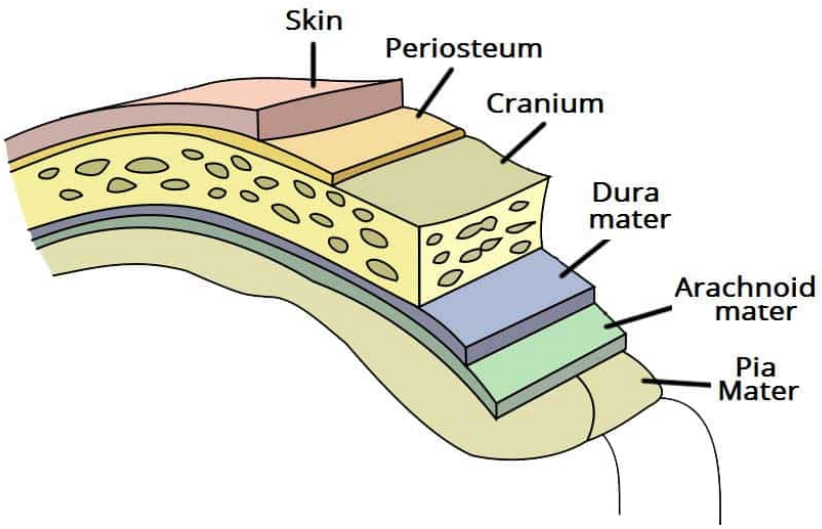

Figure 2.5: Layered structure of the meninges.

The dura mater consists of a double layer of dense irregular connective tissue and is formed by two layers, the periosteal and the meningeal [8].
Followed by the arachnoid, the intermediate layer, divided into two layers, the molecular and the trabecular. With its filaments it remains fixed to the last layer, and the pia mater, in indirect contact, due to the presence of glial cells, with the brain.
The arachnoid is made up of simple squamous epithelium and is not vascularized, while the pia mater is made up of a looser connective tissue made up of collagen and, on the contrary, is very vascularized.
The space between the arachnoid and the pia mater is called subarachnoid, and it contains the cerebrospinal fluid or liquor, produced by the choroid plexuses located in the cerebral ventricles. It has several functions, including bathing, insulating, draining and nourishing every part of the central nervous system, creating the optimal environment for cell reproduction [17].

2.8 Cerebral cortex

The brain is made up of of two types of tissue: grey matter and white matter. The grey matter is the outer layer and it has this typical colour due to the presence of cell body of neurons (soma) and dendrites. Instead, white matter is characterized by neural fibres, known as axons of neurons. The outermost layer of the grey matter is the cerebral cortex; it is from 2 to 4 mm thick and it is convoluted into several folds made of gyri and sulci because it has a large surface in a limited volume (Figure 2.6).

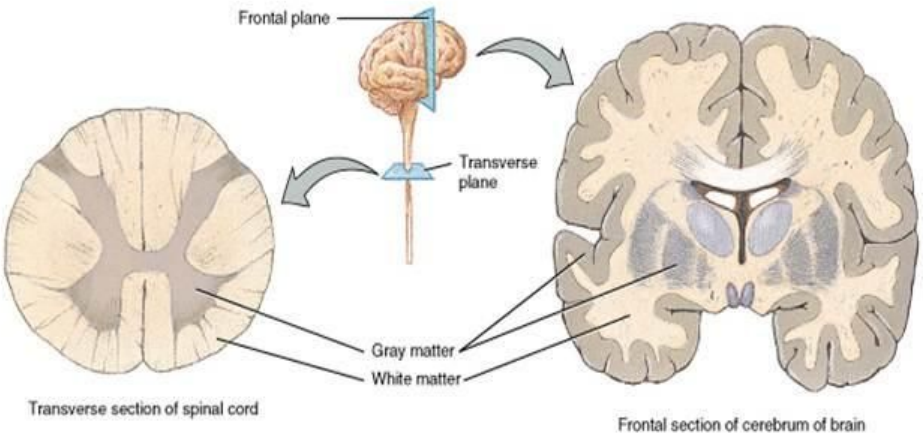

Figure 2.6: White and Grey Matter's disposition in the Brain and in the Spinal Cord.

The cerebral cortex is divided into different regions, known as projection areas and each is dedicated to specific functions. The area that manages our visual ability is located in the Occipital lobe, while the area responsible for hearing is located in the Temporal lobes. Olfactory and gustatory functions are located in the hippocampus, a section of the Temporal lobe, while more advanced thinking functions are controlled mainly by the Frontal lobe, as is the motor cortex, which is involved in planning, controlling and executing voluntary body movements. Finally, the primary somatosensory cortex located in the parietal lobe is responsible for the perception of, for example, touch, pain and temperature.

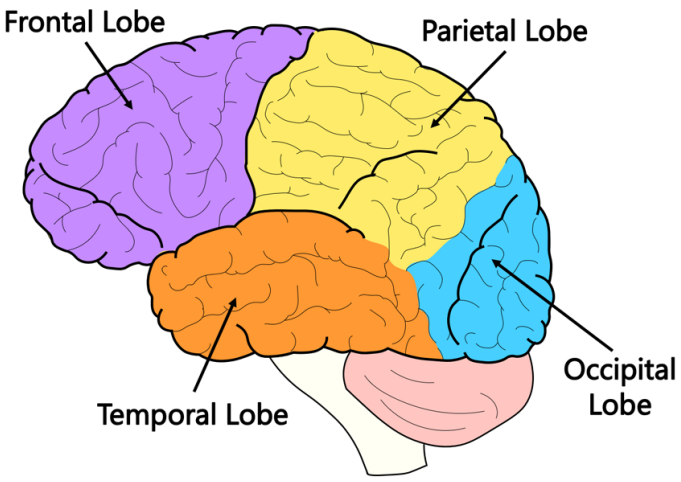

Figure 2.7: Vision of the lobes of the cerebral cortex.

In describing brain functions and projection areas, it is useful to introduce the concept of cortical homunculus of the Canadian neurosurgeon Wilder Penfield. It is a map (Figure 2.8) that depicts the areas of motor and sensory cortex associated with distinct motor functions. It is a distorted image of the body in which are represented larger the parts of the body that need more precision and a finely graded control of movement (e.g., the hand) or have greater importance for the purpose of sensory perception, are represented disproportionately large [20].

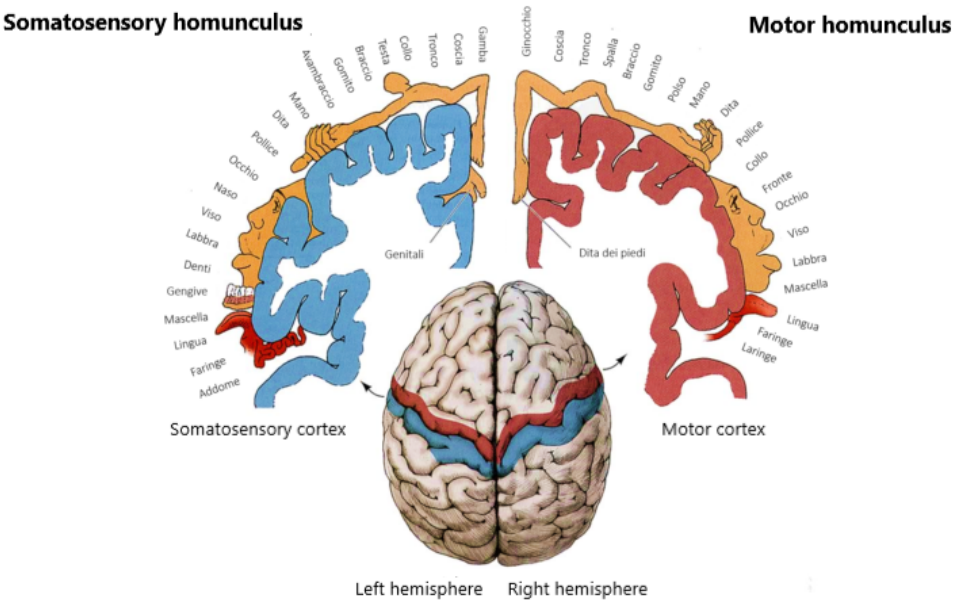

Figure 2.8: The somatosensory and motor homunculus derived by Wilder Panfield illustrating the effects of electrical stimulation of the cortex of human neurosurgical patients.

2.9 Neuron

The neuron is the basic unit of the brain and together with the neuroglia cells (more numerous than neurons and with a supporting function) and the vascular system constitute the nervous system. It contributes to the elaboration and propagation of the nervous impulse. Anatomically, it is composed of three parts: the cell body or soma, where the nucleus resides and from it branch out cytoplasmic extensions, called neurites, which are the dendrites and the axon.

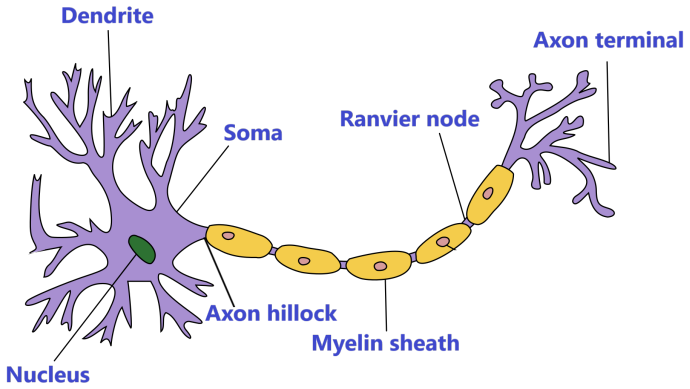

Figure 2.9: Main parts of a neuron.

The dendrites extend with a tree-like structure but but they are not responsible for the propagation of the nervous impulse as they are not good conductors and tend to decrease the speed of propagation. They, however, receive the nervous impulse from afferent cells, whereas, the axon conducts the signal towards other cells and is an excellent conductor thanks to the myelin layers (its transmission speed is approximately 150 m/s or 540

km/h). The final part of the axon is an expansion called "synaptic bouton" and through these buttons, an axon can cling to the dendrites or cell bodies of other neurons so that the nervous impulse propagates, in a chain reaction, along a neuronal network. Considering the morphology of the neuron, it can be classified based on:
1) Number and branching method of extensions:

- **Unipolar neurons**: if they present a type of extension with the value of an axon and the pyrenophore (part of the cell in which the nucleus is located) with the value of a receptor site.
- **Bipolar neurons**: if they have one axon and only one dendrite.
- **Multipolar neurons**: if they have an axon and multiple dendrites.
- **Pseudounipolar neurons**: if they appear to present only one type of ganglionic extension.

2) Formal appearance:

- **Pyramidal cells**: pyramid-shaped, the dendrites at the base are distributed horizontally, while the apical dendrite develops in height. The axon generally heads towards the innermost areas of the cortex, often entering the subcortical white matter.
- **Stellate cells**: star-shaped, also called granules, the dendrites branch in the immediate vicinity of the soma.
- **Fusiform cells**: spindle-shaped with two dendritic branches at the ends.

Based on the function it is possible to classify the neuron as:

- **Sensory neurons**: transmit information from receptors (receptor in the skin, photoreceptor in the retina...) to CNS ("afferent" neurons). They have long dendrites and short axons.
- **Motor neurons**: transmit information from CNS to effectors (muscles or glands). Have short dendrites and long axons.
- **Interneurons**: connect sensory and motor neurons within specific regions of CNS. Have short dendrites and long or short axons.

Dendrites accumulate information from other neurons or from sensory organs. This information is integrated into the axon hillock and produces an action potential (AP). It is a brief electrical signal that propagates along the axon. Information is passed through synapses (the point of connection between two neurons) and AP induces the release of neurotransmitters into the synaptic cleft and activates the next neuron. Neurotransmitters influence trans-membrane ion flow either to increase (excitatory neurotransmitters) or to decrease (inhibitory) the probability that the postsynaptic cell will produce an AP. Excitatory Post-Synaptic Potential (EPSP) and Inhibitory Post-Synaptic Potential (IPSP), are the two types of post-synaptic potential: excitatory or inhibitory depending on the type of neurotransmitter used. When there is EPSP the resting state potential becomes less negative, this phenomenon is called depolarization of the membrane, on the other hand when there is IPSP that potential becomes more negative so we call that phenomenon hyperpolarization of the membrane. If the sum of EPSP is greater than a

specific level is born an AP, while if there are a lot of IPSP or it's impossible to overcome the level, the AP can't be released. In Figure below there is a schematic representation of the AP. There is an initial phase in which the stimulus arrives, if this one is greater than the threshold the depolarization starts and also the AP. Then there is the repolarization so the membrane's potential becomes more negative and after there is a refractory period (this state is called hyperpolarization) where the neuron is inactive because of the Na+ channels. After that, the potential comes to the resting state initial potential so the neuron becomes active again.

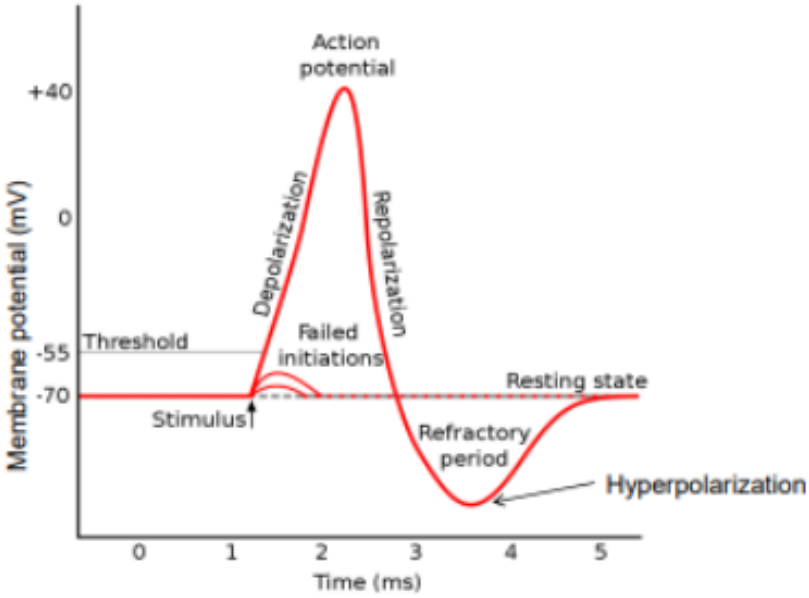

Figure 2.10: Schematic representation of the AP.

2.9.1 Neocortex layers

The neocortex is composed of six morphologically distinct layers, primarily discernible by the kinds of cells they comprise. From the most superficial to the deepest they are the following:

- I Molecular layer.
- II External granula layer.
- III External pyramidal layer.
- IV Internal granular layer.
- V Internal pyramidal layer.
- VI Polymorphic layer.

From the Figure 2.11 we can observe different types of cells in each layer of the cortex. We can divide nerve cells into two categories: pyramidal cells, the most common in the cortex, and non-pyramidal cells. The first has a pyramidal shape with an apex facing the surface of the cortex and a long axon that extends to other cortical areas or sub-cortical regions. Among nonpyramidal cells, the most numerous are the stellate cells,

with branched dendrites that originate from the cell body and terminate together with the axon, within a narrow region of the cortex. The first layer of the cortex contains very few neurons and is composed mostly of the dendrites arising from pyramidal neurons. The second consists of stellate cells and small pyramidal cells, whereas the external pyramidal layer is characterized by small to medium pyramidal neurons and it is the primary source of intracortical fibres that interconnect different areas of the cortex. The internal granular layer contains many nonpyramidal neurons and it receives much of the sensory input coming to the cortex (afferent fibres). The internal pyramidal layer is formed by the largest pyramidal cells and it is the source of efferent fibres, which carry out of the brain to various organs. The sixth layer has the greatest variety of cell types.

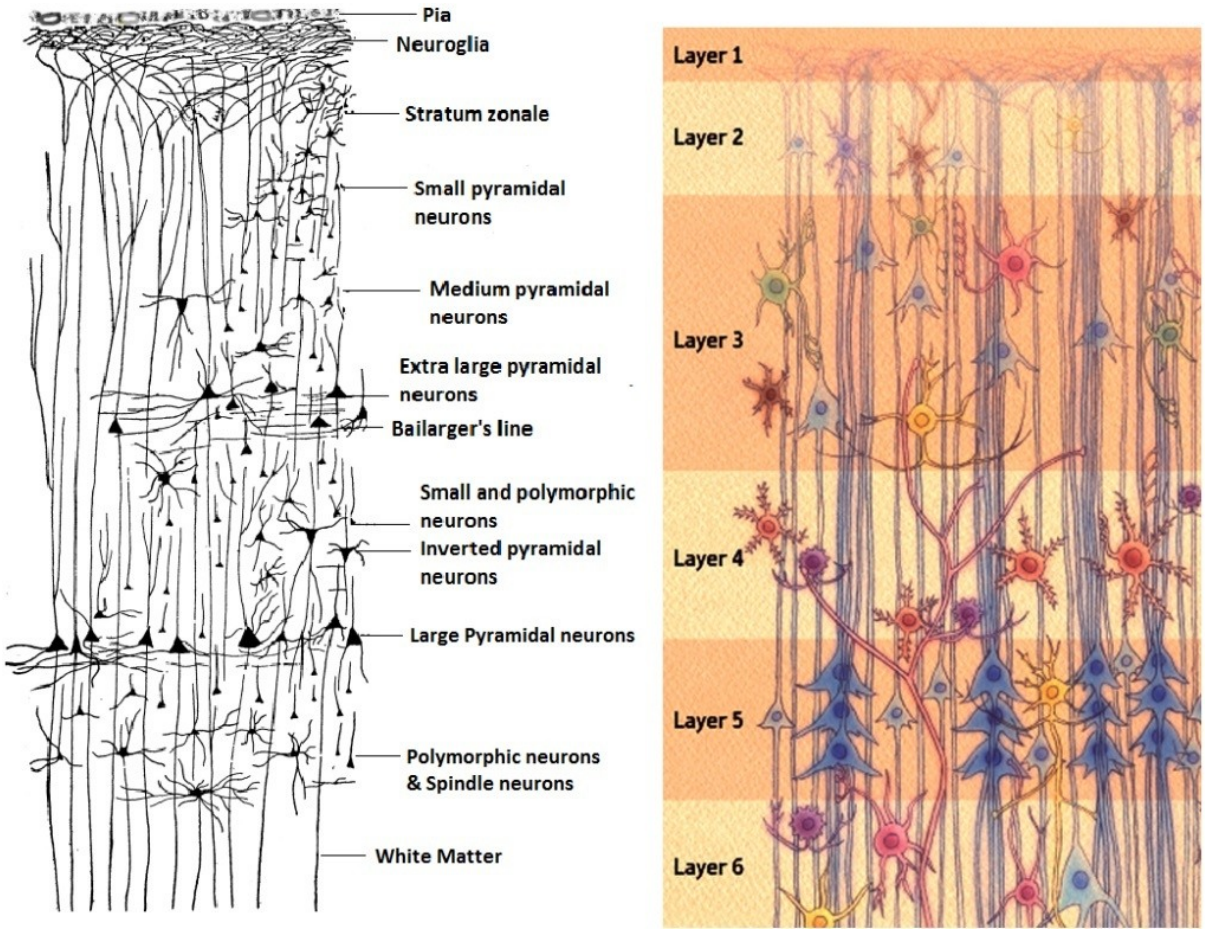

Figure 2.11: The six layers of the neocortex seen in section. At the top is the surface of the brain, at the bottom the deepest layer. On the left, the different types of neuronal cells.

2.10 Brain signal

Brain is a complex system and the brain activity extends over multiple temporal and spacial scales. Electrophysiology analyzes brain activity and involves a wide set of thechnologies to cover all these scales. We have three scales and different techniques for each ones.

- At the **microscale**, it is possible to recognize the six layers of neocortex and we it possible to obtain intracellular voltage recordings with direct measurements. It is

a powerful, but tedious method and for this reason it is limited to a few neurons per experiment. Patch clamp or sharp microelectrode techniques are used. Patch clamp is a method to measure the currents that pass through individual ion channels present in the cell membrane. Sharp microelectrode is a needle electrode inserted inside the cell, so that the membrane potential can be measured.

- At the **mesoscale** it is possible to study a population of neurons with extracellular recording techniques. It is obtained a signal that is formed by simultaneous and long-term recordings of Local Field Potentials (LFPs) and Extracellular APs at ms time scale. These could be done by Microelectrode Arrays (MEAs) and polytrodes. The first are arrays of microelectrodes while the latter are instruments composed of multiple electrodes arranged in a single structure, often aligned along a probe.

- At the **macroscale** it is possibile to study large areas of the brain with EEG (electroencephalography) that detects spontaneous and evoked electrical activity from the scalp with low spatial resolution (cm range). The main techniques of acquisition are:

  - **Electrocorticography (ECoG)** recording the electrical activity of the cerebral cortex using electrodes placed directly on the surface of the brain. Its use involves monitoring brain activity during surgery and to map cortical functions in patients with epilepsy.

  - **Intracranial Electrocorticography (iEEG)** is a technique of directly recording the electrical activity of the brain using surgically implanted electrodes. It is primarily useful for the presurgical evaluation of patients with drug-resistant epilepsy.

  - **Functional Magnetic Resonance Imaging** (fMRI) is a noninvasive imaging technique that measures brain activity by detecting changes in blood flow acquiring BOLD signal (Blood Oxygen Level Dependent). It is used to map brain functions, study functional connectivity and identify brain regions activated during specific tasks.

  - **Positron Emission Tomography** (PET) uses radioactive tracers that produce gamma rays to visualize metabolic processes in the body. It is used to study brain metabolism, blood perfusion, and to diagnose diseases such as cancer and neurodegenerative diseases.

  - **Magnetoencephalography (MEG)** measures the magnetic fields generated by the neuronal activity of the brain. Used to map brain activity, study cognitive functions, and localize areas prone to epileptic seizures.

In Figure 2.12 we can observe in (A) the different techniques for acquiring brain signals belonging to the macroscale, both invasive and non-invasive.
Instead, in (B) the two intracellular and extracellular recording techniques belonging to the microscale and mesoscale respectively are represented [19].

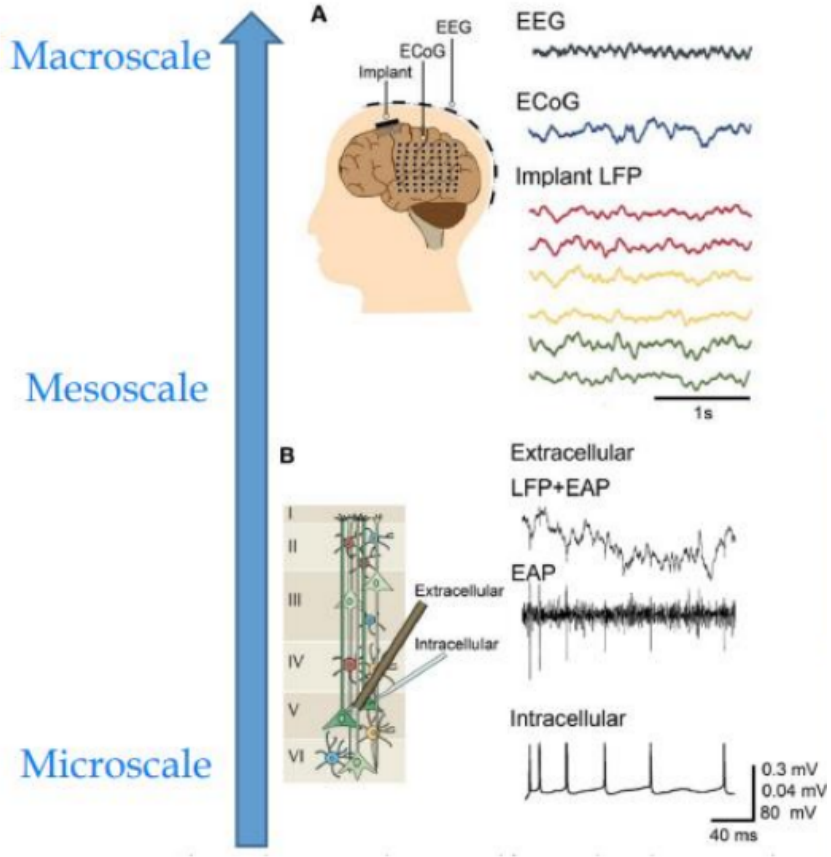

Figure 2.12: Three levels of acquisition of brain signals and different techniquesb [19].

2.10.1 Electroencephalography(EEG)

Electroencephalography is a diagnostic test that measures the electrical activity of the brain. In particular, it records variations in the electric field generated by groups of pyramidal neurons.
The depolarization that occurs along the axon is not measured, but only the post-synaptic potentials.
The synapse is a structure specialized in communication between cells, which leads to the transfer of electrical potential towards the post-synaptic membrane. A chemical transmission occurs in the synapses and in this area it is possible to measure a potential that has greater amplitude and is slower. The EEG recording is a noninvasive technique, relatively easy to use and it not require any surgery, as it can also be implemented in a non-hospital environment.
It has several utility as diagnosis of epileptic seizures, monitoring of anesthesia during surgery, evaluation of brain lesions, study of sleep disorders, evaluation of neurological and psychiatric disorders, such as Alzheimer's disease, Parkinson's disease, schizophrenia and depression, monitoring of patients' brain activity in intensive care units and in the application of therapies such as deep brain stimulation.
The first EEG recording machine was introduces to the world by Hans Berger in 1929. Berger, a neuropsychiatrist of Jena University in Germany, used the German term "elektrenkephalogramm" to describe the graphical representations of electrical currents generated in the brain. The Berger's device consisted of a "string galvanometer" that used

the physical principle of a magnetic coil, which upon the passage of current, creates a vibration of a quartz fiber [23]. To make the acquisition permanent, the vibrations were reflected on a strip of photographic paper moving at a constant velocity in a light-tight chamber. He suggested that the brain currents change depending on the functional state of the brain, such as sleep, anesthesia and epilepsy. In fact the EEG signal is highly correlated with the subject's level of consciousness. If brain activity increases, the EEG signal shows a higher main frequency and a lower amplitude. This was a revolutionary idea that helped to create the new brench of medical science called neurophysiology.

2.10.2 EEG characteristcs

it is possible to identify various parameters and characteristics on the EEG signal:

- **Amplitude**: variable between 10 and 500, can be distinguished in normal EEG, low ($< 30\ \mu$V), medium (30 μV-70 μV) and high ($> 70\ \mu$V).
- **Morphology**: with which a repetitive signal of a given frequency manifests itself, this can be described as "polymorphic" or "monomorphic".
  - By polymorphism we mean the succession of potentials belonging to the same frequency band, but with non-regular periodicity and amplitude often different from one component to another.
  - By monomorphism we mean the regular succession of potentials having exactly the same frequency and, often also the same amplitude.
- **Topography**: the definition of the brain areas in which an electrical event occurs. It is identified by referring to the classic anatomical distinction of the cerebral hemispheres into frontal, parietal, occipital and temporal, right and left lobes.
- **Symmetry/Asymmetry**:
  - those signals are symmetrical which, appearing on both hemispheres (even if at successive times), have the same characteristics of frequency, amplitude and duration on both sides.
  - those events that occur only in one hemisphere or, if bilateral, have different characteristics on the two sides are asymmetrical.
- **Synchrony/Asynchrony**: synchrony concerns the moment of appearance of certain electroencephalographic events and they are defined:
  - Synchronous those that occur simultaneously on both sides.
  - Asynchronous those that occur on the two hemispheres at different times.

2.10.3 Components of EEG signals

EEG is a complex signal that includes many frequency components. They are called also "EEG rythms" and are used principally in sleep analysis because each band is associated with a cognitive state of the subject. The main frequency components of the EEG signal and their main characteristics will be reported in Table 2.1 below. Instead, in Figure 2.13 we can observe the waveforms of the frequency components of the EEG signal.

Type of rhythm	Frequency (Hz)	Amplitude (µV)	Associated status
delta (δ)	0.5-3	20-200	In pathological conditions such as epilepsy, during anesthesia, or in very deep sleep
theta (θ)	3-7	5-100	State of mental relaxation or altered state of consciousness
alpha (α)	8-13	10-200	State of wakefulness but at mental rest
beta (β)	14-30	1-20	Attention, concentration
gamma (γ)	> 30	1-20	Associated with learning process, memory and higher consciousness

Table 2.1: EEG rhythms and associated states.

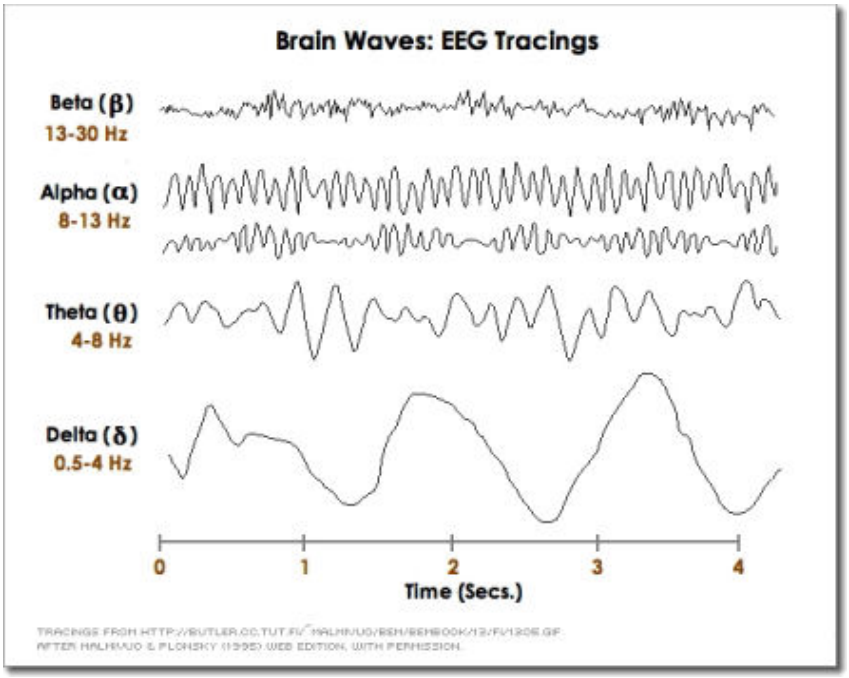

Figure 2.13: Waveforms of the EEG rythms.

2.10.4 Mu waves

Mu waves have relevance both in the development of the brain computer interface and in neurorehabilitation therapies. These are waves with frequencies between 7 and 12.5 Hz (very similar to alpha waves). These waves are present in the motor cortex, but when one carries out a voluntary movement or with the intention of carrying it out they disappear. They are easy to use because they have very specific in-band and topological localization. Amputee patients, with spinal injuries, with stroke outcomes and with dystrophy could benefit from these devices.

2.10.5 EEG recordings and International 10-20 System

There are two approaches for EEG recording: it can be done non-invasively by situating electrodes on the scalp, or invasively through methods such as electrocorticography (ECoG), where electrodes are implanted directly into the cerebral cortex.
Here, we will analyze the sampling of the EEG signal with the non-invasive technique. This technique consists of placing electrode directly on the scalp of the subject; depending on the type of electrode used, it is placed directly in contact with the skin (dry electrode), or a conductive is placed between the skin and the surface of electrode, with the purpose of lowering the impedence of contact (wet electrodes).The last are considered the "gold standard" because are the most used and usually consist of one combination of silver and silver chloride (Ag/AgCl).

The International Federation's 10-20 system is an internationally recognized method, first presented at the Fourth International EEG Congress in Brussels in 1957 by Herbert Jasper, to standardize the method of EEG placement as it ensures coverage of all regions of the brain.
First, the distance along the sagittal plane of the subject is calculated between two anatomical landmarks: "inion" (prominence at the base of the occipital bone) and "nasion" (upper nose junction). Also, starting from nasion and moving 10% of the preliminarly calculated I-N distance (axis dividing the cranium into right and left hemisphere), find the position of the fronto-polar (FP) electrode and moving along the meridian intersecting that point, place one electrode on the right and one on the left moving 10%. On the same z-line (sagittal axis), moving in steps of 20% you place the frontal electrode (F), also you find the middle one (C), the parietal one (P) and, at the end, the level of the right and left occipitals (+20%) (O1 and O2), with which you arrive at 90% of the nasion-inion distance, at a distance of 10% from the inion. From each point linked on the z-line (Fz, Cz and Pz) we move an additional 20% along the meridian identified, to position the other lateral electrodes. Odd numbers identify the left hemisphere, while indeed numbers identify the right hemisphere.

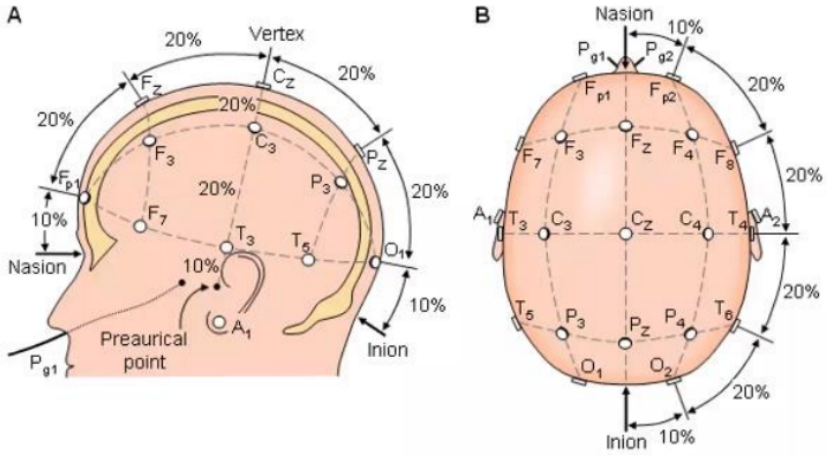

Figure 2.14: Lateral (A) and dorsal (B) view of electrode placement on the scalp according to Standard 10-20.

To make indeed more precise measures, the so-called "extended 10-20 extended," in which others electrodes are inserted at positions intermediate between those provided by the traditional 10-20 system; in this way, up to further than 70 possible positions for electrode placement. In the figure below (Figure 2.15), using color coding it is possible to observe the various areas of the brain covered by the electrodes.

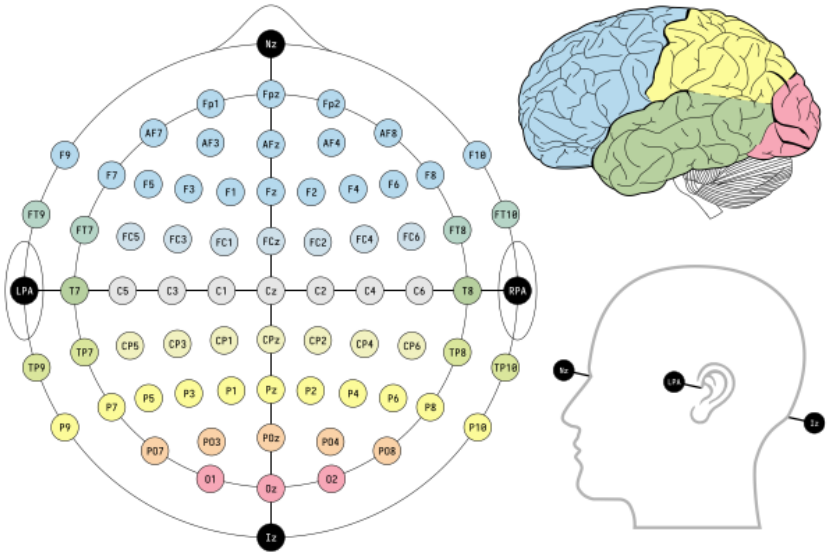

Figure 2.15: Positioning of the 73 electrodes on the scalp according to the Extended 10-20 Standard.

2.10.6 Mono-polar vs bipolar acquisition

From the point of view of EEG signal acquisition, once the electrodes have been deposited, it is necessary to choose the acquisition mode between mono-polar, i.e. the potential seen by the exploring electrode compared to the reference electrode chosen for the setup, or bipolar mode where, rather, the read potential is the difference between the potentials read by adjacent individual scanning electrodes. The difference lies above all in the withdrawal volume of the acquisition. In fact, if the bipolar mode is used there is a

difference between two different electrodes, reducing the contribution from the more distant sources of the electrodes, and maintaining the contribution of the sources closest to the electrodes. Depending on the study you want to do, it may be more appropriate to use one system rather than the other. In the figure below we have a bipolar sampling in A, thus we see 4 electrodes and 3 signals given by the differences 1-2, 2-3 and 3-4, and in B a mono-polar sampling. In the case of bipolar sampling, it is understood that the source of the potential is located between electrodes 2 and 3 because they both read the same signal and therefore the difference between the two will be an almost flat signal. In the case of mono-polar sampling, it is possible to detect the source by observing the amplitudes of the signals, as you will be very close to the electrode that captures the largest signal.

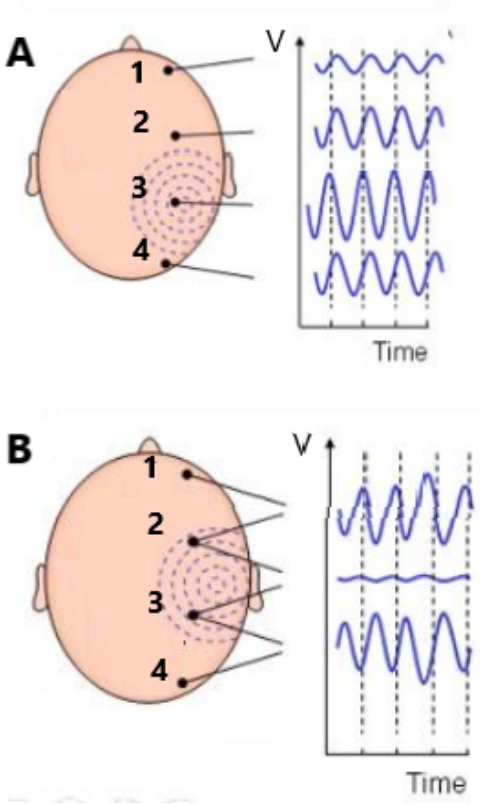

Figure 2.16: Example of mono-polar acquisition (A) and bipolar acquisition (B).

2.10.7 EEG artifacts

Although the electroencephalograph is designed to record brain activity, one of the main problems it suffers from is the contemporaneous detection of signals of different origins, called artefacts, coming from other sources of an extra-cerebral nature. Artifacts are disturbances (or noises) that contaminate the signal of interest, making it delicate to determine and extract useful information of brain origin.
Based on their origin, artifacts can be classified into two classes: those of **physiological origin** and those of **extra-physiological origin**. Artifacts of physiological origin come from biological sources internal to the body, other than the brain, while those of extra-physiological origin derive from sources external to the body, similar as the equipment used for recording or the surrounding environment. By using appropriate ways, it is possible to reduce or eliminate the artifacts present in the signal, therefore obtaining a

clean signal that allows useful information to be recognized more clearly and accurately.

The main artifacts of physiological origin are:

- **Eyeblinks** are among the most frequent artifacts and manifest as negative waves of very high amplitude in bi-frontal areas. This happens due to the Bell phenomenon. The cornea, which has a positive charge, and the retina, which has a negative charge, influence the signal: when you blink, your eyes lift slightly, bringing the cornea closer to the frontal electrodes Fp1 and Fp2. As a result, these electrodes detect a positive signal that is reflected in the EEG. When the subject closes his eyes, the positive pole approaches the frontal electrodes, causing downward deviations that are symmetrical and short-lived. As soon as he reopens his eyes, the positive pole moves away from the frontal electrodes, producing an upward deflection.

- **Lateral eye movements**: Positive or negative waves appear on channels F8 and F7. This is due, once again, to the positive charge of the cornea and the negative charge of the retina. In particular, if the movement is to the right, the wave is positive on F8 e negative on F7. When looking to the right, the right cornea approaches the F8 electrode, which sees a positive charge; the left retina approaches electrode F7, which therefore sees a negative charge.

- The **ECG artifact** is characterized by waveforms synchronized with the QRS complex of the electrocardiogram. They generally occur mainly or exclusively on the left side, since the heart is located in the left side of the chest, and have a relatively low amplitude. This artifact is apparent in EEG signals as a small wave peak corresponding to the QRS complex, and is generally more visible in posterior channels, similar as the occipital ones.

- **Myogenic potentials** are among the most common artifacts. They can generally be easily identified thanks to their morphology and duration. They are characterized by high-frequency, often low-amplitude activity that overlaps with normal brain rhythms, is usually most prominent in the waking state, and, furthermore, often occurs in clusters.

- **Chewing and tongue artifact (hypoglossal)** are quite easy to notice on the EEG. The masticatory artifact is actually just a muscular artifact of the temporalis muscle and is characterized by sudden onsets, intermittent bursts of very fast generalized activity (muscular artifact). It is easily identifiable through video studies, as it is sufficient to watch the video to correlate it, but even without video studies the masticatory artifact usually does not have a common morphology with any other important physiological activity.

- **Skin artifact**: Sweat and other common physiological processes are capable of change the impedance of the electrodes. This causes low-frequency (typically less than 0.5 Hz) and relatively low-amplitude drifts of potentials and waves that occur because the sodium chloride in sweat carries a charge, which is detected by the EEG electrodes. But this artefact can also be created due to the presence of a hematoma in the area under the electrode, in which case the signals are characterized by a general attenuation. The sweat artifact does not follow a precise pattern in terms of localization and can be bilateral, unilateral or even focal for small number of electrodes.

The main extra-physiological artefacts are:

- **Network artifact**: interference at 50 Hz (in Europe and elsewhere), or at 60 Hz (in the United States) due to poor electrode-skin contact. The electrical artifact is a very fast and monotonous activity, and it is possible to use the notch filter to remove it selectively, furthermore this removed frequency content will not affect the interpretation of the EEG signal as this frequency value is not related to no brain activity.

- **Movement of the electrical cables**: during acquisition, care must be taken to ensure that the cables are not dangling or under tension, because their movement is add to the EEG signal creating an artifact;

- **High electrode-skin impedance**: artefact due to poor contact between the electrode and the skin. To avoid such trouble first. During the acquisition of EEG signals, the electrode-skin impedance is measured which normally should not be less than 5 kOhm if wet contact electrodes are used.

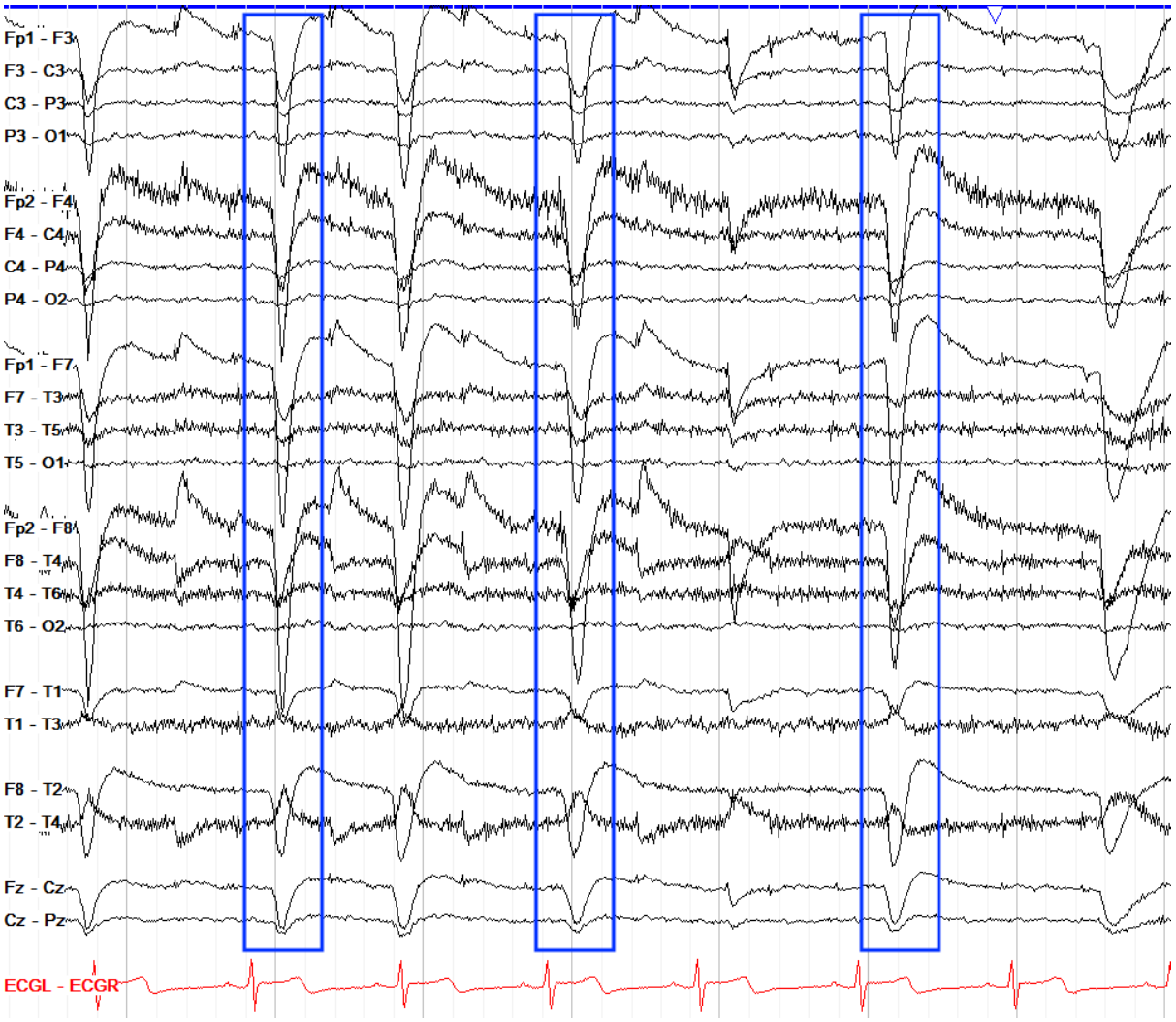

Figure 2.17: Example of eyeblink artifact.

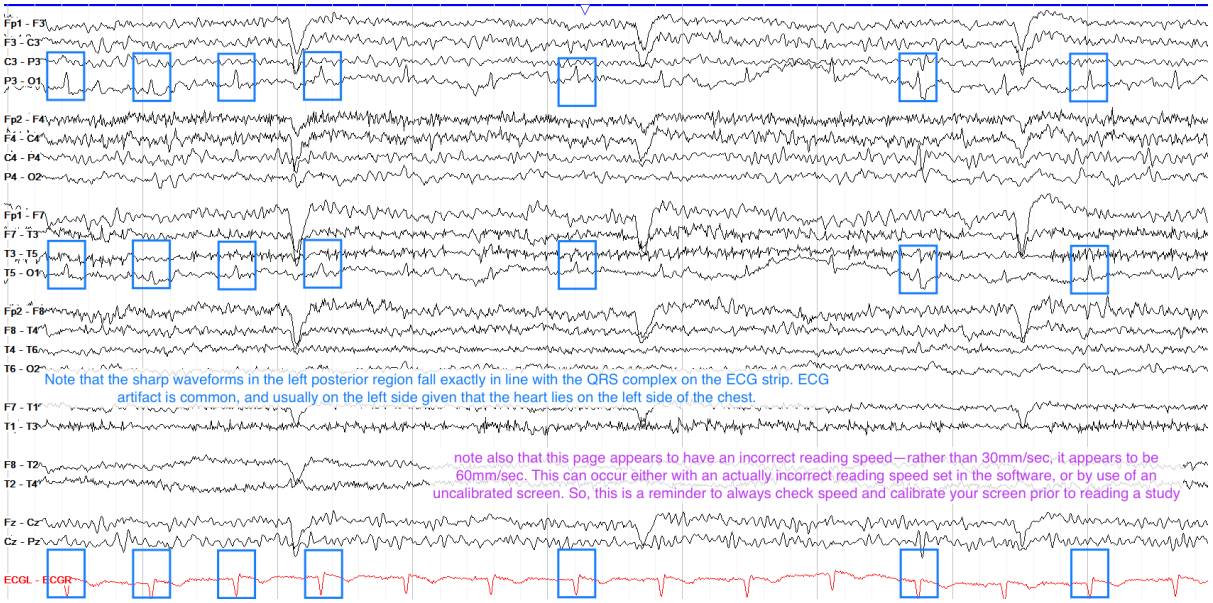

Figure 2.18: Example of cardiac artifact.

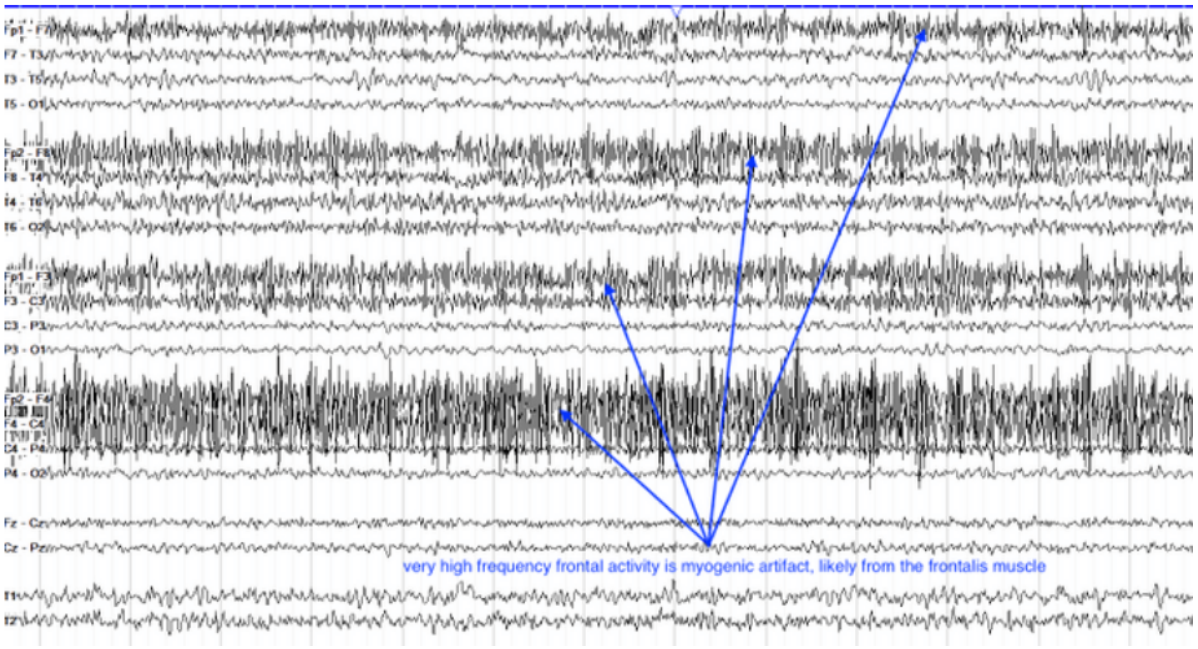

Figure 2.19: Example of Myogenic artifacts.

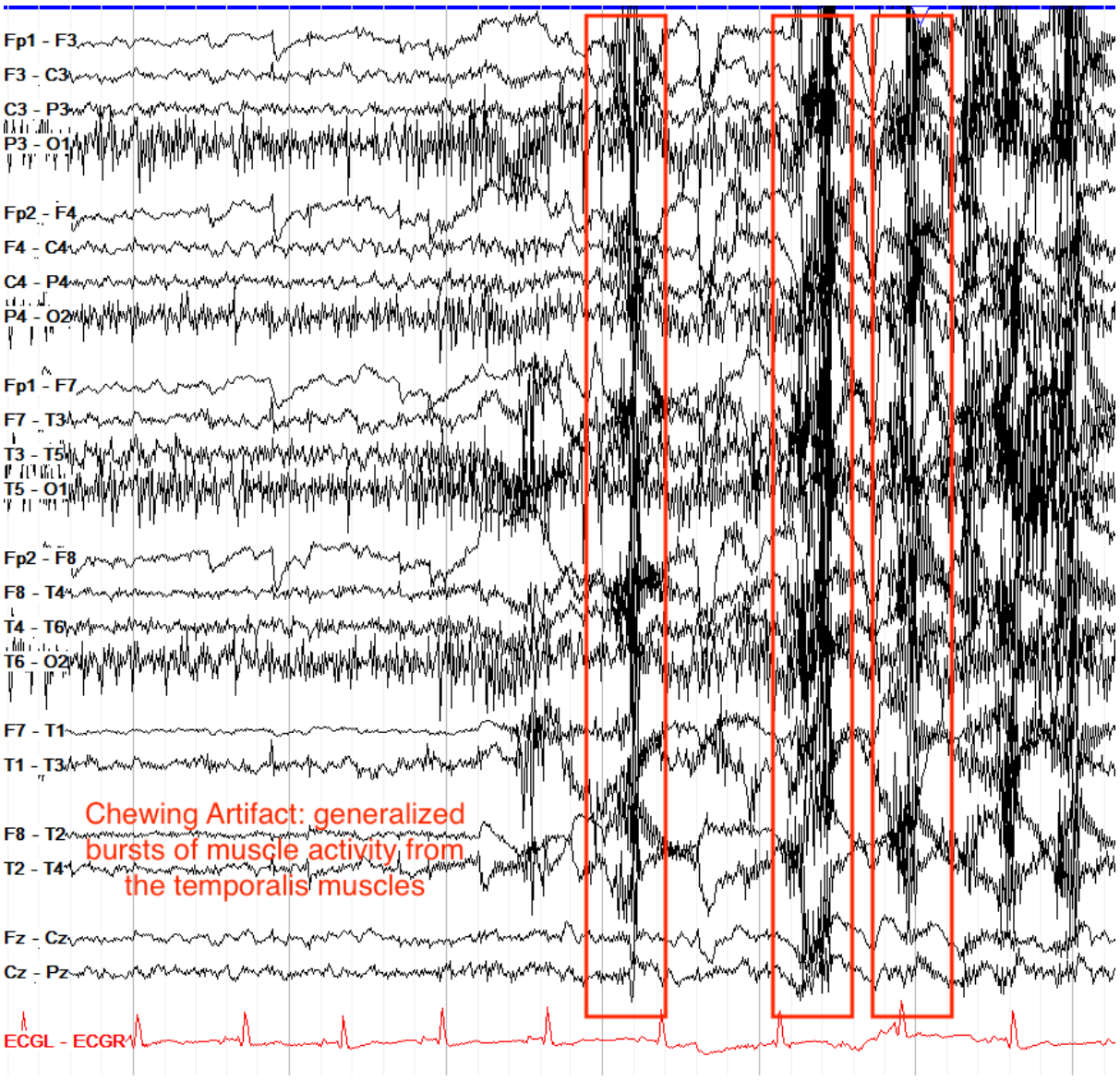

Figure 2.20: Example of chewing and tongue artifact.

2.11 BCI

A brain-computer interface (BCI) is a hardware and software communications system that allows cerebral activity just to control computers or external devices. It creates a new non-muscular channel for giving a person's intentions to external devices [18]. The main purpose of BCI research is to provide communications capabilities to disabled people who are totally paralyzed or 'locked in' by neurological neuromuscular disorders, such as sclerosis, brain stem stroke, or spinal cord injury. Brain-computer interfaces use brain signals to gather information on user intentions. To achieve this, BCIs require a recording stage that measures brain activity and translates this information into electrical signals. There are two types of brain activity that can be monitored: electrophysiological and hemodynamic.
Electrophysiological activity is generated by electrochemical transmitters exchanging information between neurons, which produce ionic currents.
This activity is measured using electroencephalography (EEG), electrocorticography (ECoG), magneto-encephalography (MEG), and the acquisition of electrical signals in single neurons.
Hemodynamic response, on the other hand, is a process where blood releases glucose to active neurons at a higher rate than to inactive neurons. The delivered glucose and oxygen result in a surplus of oxyhemoglobin in the active area and a change in the local ratio of oxyhemoglobin to deoxyhemoglobin. These changes can be quantified using neuroimaging methods such as functional magnetic resonance imaging (fMRI) and near-infrared spectroscopy (NIRS).
These methods are considered indirect because they measure the hemodynamic response, which, unlike electrophysiological activity, is not directly related to neuronal activity.
Most current BCIs obtain relevant information from brain activity through electro - encephalography due to its high temporal resolution, relatively low cost, high portability, and minimal risks to users. BCIs based on EEG consist of a set of sensors that acquire EEG signals from different brain areas.
Non-invasive approaches have been used by severely and partially paralyzed patients to regain basic forms of communication and to control neuroprostheses and wheelchairs. These approaches are based on an endogenous method, which involves the self-regulation of brain rhythms and potentials without external stimuli, achieved through neurofeedback. Additionally, these methods are bidirectional, as they include both recording and stimulating.
Invasive recording techniques as electro - corticography or intracortical neuron recording are used to control neuroprostheses with multiple degrees of freedom. Invasive modalities require the implantation of microelectrode arrays inside the skull, which involves significant health risks. In BCI research, there are two invasive modalities: electro - corticography, which places electrodes on the surface of the cortex, and intracortical neuron recording, which implants electrodes inside the cortex.
These methods are based on an exogenous approach that uses neuronal activity in response to an external stimulus and are unidirectional [14].
The different steps that form a standard BCI as can be seen from:
1) signal acquisition: Electrical signals reflecting brain activity are acquired from the scalp, from the cortical surface, or from within the brain.
2) preprocessing or signal enhancement: remove the artifacts in the control signals and improve the performance.

3) feature extraction and classification: The signals are analyzed to extract signal features (such as amplitudes of EEG rhythms or firing rates of single neurons) that reflect the user's intent.
4) the control interface: Features are translated into commands that operate application devices that replace, restore, enhance, integrate, or improve natural resources [24].

2.11.1 Control Signal Types in BCI

A BCI interprets user intentions by monitoring cerebral activity. Brain signals related to cognitive tasks involve various simultaneous phenomena that can serve as control signals for BCIs, such as visual evoked potentials, slow cortical potentials, P300 evoked potentials, and sensorimotor rhythms. Event-related potentials (ERPs) in the EEG are manifestations of neural activity triggered by specific events and can be extracted by averaging multiple trials of similar events. ERPs include exogenous components (primary sensory, < 150 ms) and endogenous components (information-processing activity, longer latency).

Visual Evoked Potentials (VEPs)

VEPs are modulations of brain activity in the visual cortex following visual stimuli. Their amplitude increases significantly when the stimulus moves closer to the central visual field. VEPs can be classified by the morphology of optical stimuli, the frequency of visual stimulation, and the area of stimulation. There are transient VEPs (TVEPs) occurring at stimulation frequencies below 6 Hz and steady-state VEPs (SSVEPs) at higher frequencies. SSVEPs are less susceptible to artifacts from eye movements and electromyographic noise and allow users to select a target by gazing at it. An f-VEP-based BCI uses unique stimulation frequencies for each target, requiring minimal training but is not suitable for patients with severe visual or motor limitations.

Slow Cortical Potentials (SCPs)

SCPs are slow voltage changes in the EEG associated with changes in cortical activity. Negative SCPs indicate increased neuronal activity, while positive SCPs indicate decreased activity. SCPs can be self-regulated by both healthy users and paralyzed patients to control external devices via a BCI. However, SCP-based BCIs provide relatively low information rates and require long training periods and continuous practice.

P300 Evoked Potentials

P300 evoked potentials are positive EEG peaks caused by infrequent stimuli. They appear around 300 ms after attending to a rare stimulus, with the amplitude being larger for less probable stimuli. P300-based BCIs require no training and are useful for individuals with limited eye movements, such as ALS patients. However, performance may decline as users become accustomed to the infrequent stimuli.

Sensorimotor Rhythms (mu and beta rhythms)

Sensorimotor rhythms, including mu (7–13 Hz) and beta (13–30 Hz) rhythms, vary in amplitude during motor tasks. These rhythms can be used to control BCIs because people can learn to voluntarily modulate them. Amplitude modulation, known as event-related desynchronization (ERD) and event-related synchronization (ERS), is generated by sensory stimulation, motor behavior, and motor imagery. This allows the use of sensorimotor rhythms for endogenous BCIs, which do not require actual movement but rather imagined movement to produce similar ERD and ERS patterns.

2.12 The enigma of Consciousness

Consciousness is one of the most fascinating and enigmatic mysteries of the human experience. Its nature has led to endless analyses, explanations and debates by philosophers, theologians and scientists since the time of the Greeks with Aristotle. In some explanations it is synonymous with mind and other times an aspect of mind. In the past it was "soul", the "inner life", the world of introspection, private thought, imagination and volition [16]. Nowadays, consciousness relates to our awareness of our thoughts, emotions, feelings, perceptions and surroundings environment at any one moment in time. It shapes our perception of reality(what we believe to be real) and our sens of selfs. It may or may not be ever-evolving awareness, awareness of awareness, or self-awareness. A significant part of scientific research on consciousness is characterized by studies with the aim of investigating the relationship between the experiences reported by the subjects under study and the brain activity that occurs simultaneously. These studies, known as studies on the neural correlates of consciousness, aim to identify that specific activity in a certain area of the brain or in a particular pattern of global brain activity that can act as an indicator and make it clear that the subject has conscious awareness. To measure brain activity in such studies, various brain imaging techniques, such as EEG and fMRI, have been employed.

2.12.1 What is Consciousness?

The term consciousness is often used interchangeably with the term mind, but they are two different concepts. Mind is the marvelous space where are situated perceptions, reasonings, where happen the translations of perceptions, for example,of objects and actions in words. In his article, Andrej Ule defines mind as "human ability and activity to consciously grasp and understand specific contents and objects of human activity" and and connects it to the concepts of "intellectus", associated with a type of intuitive understanding and "ratio", associated with rational and logical thinking.
Antonio Damasio, director of the Brain and Creativity Institute at the University of Southern California, where he is professor of neurology, neuroscience and psychology, proposes that consciousness derives from the interactions between the brain, the body, and the environment. According to this theory, consciousness is not a unitary experience, but rather emerges from the dynamic interaction between different brain regions and their corresponding bodily states [7]. In his seminar "Exploring the Enigma of Consciousness: Artificial Intelligence and the Mind" [3], A.D. argues that consciousness is the experience of thoughts and life that occurs within our organism. This experience is subjective and represents a unique perspective, which is usually the perspective of our self. The key idea is that this perspective or consciousness is created by an asymmetric construction within our brain and organism, which generates the self's experiences and perspectives on everything that happens in our mind.

2.12.2 States of consciousness

Scientist and philosophers have long tried to explain how th brain generates conscious experiences. Some question whether the objective methods of science can ever fully understand a phenomenon as subjective as consciousness. However, researchers have started to identify the shifts in brain activity associated with awareness and have also

developed some intriguing theories about the origins and evolution of consciousness. How the brain evokes conscious awareness from the electrical activity of billions of individual nerve cells remains one of the great unanswered questions. In the 1990s, philosopher David Chalmers, described the impossibility to demonstrate somebody else that Each of us knows that we are aware, in terms of thoughts, perceptions and feelings. This is identified as the "hard problem" of consciousness.

Three dimensions of consciousness Consciousness includes distinct dimensions that can be measured. Among the three most important are:

- Awake or state of physiological activation.
- Awareness, or the ability to experience conscious mental experiences, including perceptions, thoughts, and emotions.
- Sensory organization, or how different perceptions and more abstract concepts intertwine to create a unified conscious experience.

These three dimensions interact dynamically to determine our overall state of consciousness at any given moment. For example, during full wakefulness, we experience increased awareness, but as we fall asleep during the night, both wakefulness and awareness decrease.

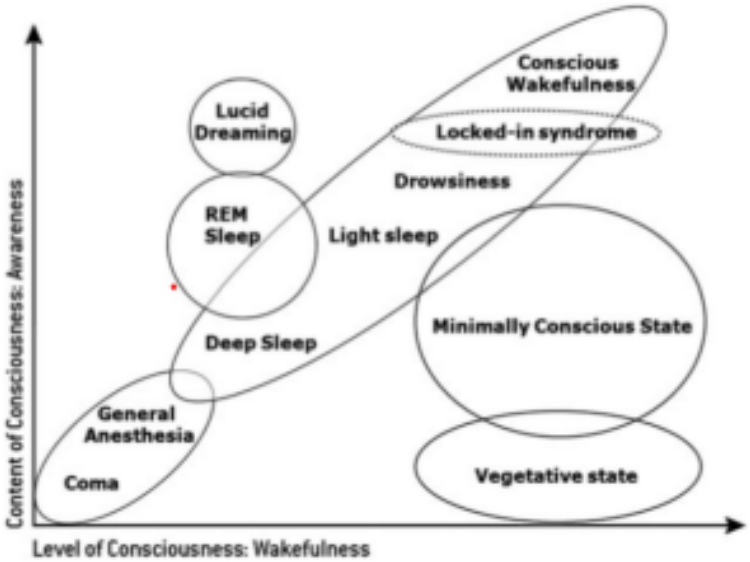

Figure 2.21: Interaction of the two major components of consciousness, wakefulness and awareness, in different states of consciousness.On the X-axis is the level of wakefulness and on the Y-axis is the content of awareness [14].

2.12.3 Disorders of consciousness

Medical conditions that repress consciousness are referred to as disorders of consciousness. These conditions are often described as any deviation from full self-awareness, ranging from a reduction in self-awareness and arousal to a complete absence of self-awareness. These conditions can arise following severe brain injuries, head trauma, stroke or other brain pathologies and neurodegenerative diseases.

Minimally conscious state
The minimally conscious state is a state of altered consciousness in which a person shows signs of reduced awareness and responsiveness, but less pronouncedly than a fully conscious person. This state represents a higher level of consciousness than coma and the vegetative state, but still less than complete awareness. People in a minimally conscious state may exhibit limited behaviors or responses, such as opening their eyes, following objects with their gaze, make sounds or respond in a very basic way to verbal or physical commands. However, these responses are often inconsistent and do not indicate full awareness of the self or surroundings.

Vegetative State
The vegetative state, also known as persistent unconsciousness, is a serious and complex condition in which a person appears to be awake, but lacks awareness of himself and his surroundings, is characterized by a significant loss of cognitive and conscious functions. In the setting of a vegetative state, the patient may have cycles of sleep and wakefulness, may open his eyes, and may perform reflex movements or involuntary behaviors, but shows no signs of intentional consciousness or significant interaction with the environment. For example, he may make sounds in response to stimuli, but does not demonstrate awareness or complex communication skills.

Coma
Coma is an altered state of consciousness characterized by a profound and prolonged loss of consciousness, in which the patient does not respond to external stimuli and is unable to communicate or interact with the surrounding environment. Coma is an emergency medical situation and requires immediate attention from a specialized medical team. Doctors try to identify the underlying cause of coma through brain tests and imaging. Treatment given after the onset of coma aims to stabilize the patient, manage underlying conditions and protect the brain from further damage. The prognosis for patients in a coma varies greatly based on the severity of the underlying cause and the effectiveness of treatment. Some patients may emerge from a coma and recover fully, while others may remain in a long-term state of altered consciousness. In some cases, coma can progress to a vegetative state or a minimally conscious state.

Brain death Brain death is a medical condition defined as the complete and irreversible loss of all brain function, including brainstem function. In this situation, the brain ceases to function permanently, and the patient is clinically dead, although the heart and lungs can be kept functioning artificially via life support devices. The diagnosis of brain death requires an extremely rigorous and precise medical evaluation. Typically, doctors use specific clinical criteria, including neurological tests and observation of brain reactions to pain and visual stimuli, to determine brain death.

Locked-in syndrome Locked-in syndrome is a rare neurological condition in which a person is fully conscious, but paralyzed and unable to move or communicate through traditional means, such as speech or body movement. This state of paralysis is caused by specific lesions or damage in the brainstem, which is the part of the brain responsible for controlling body movements. Despite total paralysis, people with "locked-in" syndrome retain their full consciousness and cognition. Often they are only able to move their eyes or

eyelids. Communication may be possible through the control of assistive communication devices, for example with the help of Brain Computer Interface [14].

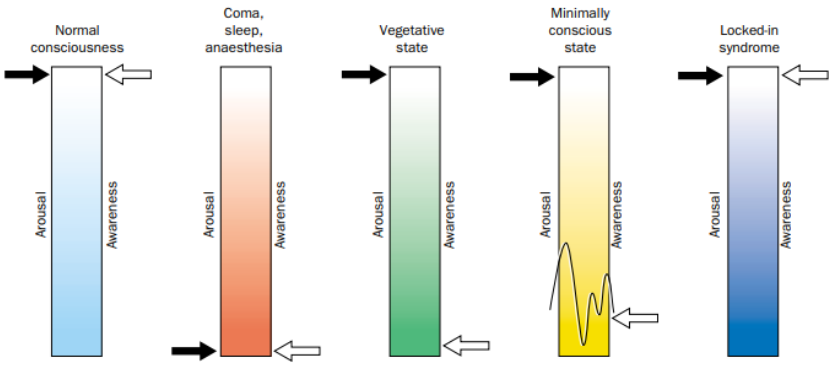

Figure 2.22: Comparison of levels of arousal and awereness of different states of disorder of consciousness compared to the state of normal consciousness.

2.12.4 Scales for assessing the state of consciousness

The scales for evaluating states of consciousness are some of the tools used in the medical field to evaluate the level of consciousness of an individual, particularly in cases of brain lesions or alterations in the state of consciousness. These scales help health professionals assess the level of impaired consciousness and monitor changes over time.

Glasgow Coma Scale (GCS)
The GCS is one of the most widely used scales of disorders of consciousness. It evaluates three main aspects of the level of consciousness: eye opening, verbal response and motor response. Each aspect is rated on a scale from 1 to 4 or from 1 to 6, depending on the version used. The total score is expressed synthetically with a number which is the sum of the ratings of each individual function and can vary from 3 (the worst) at 15 (the best). The GCS is used to evaluate patients with brain injuries, strokes and other medical conditions that can affect consciousness. The Glasgow scale is less suitable for children, especially under 36 months of age, when the young patient has not yet developed mastery of language: this is why the Pediatric Glasgow Coma Scale was eveloped [11] .

Reaction Level Scale (RLS)
The Reaction Level Scale is another scale used to evaluate the level of consciousness and evaluates the patient based on the ability to respond to verbal and physical stimuli, also evaluating the presence of reflexes. The RLS has a scale ranging from 1 (worst) to 8 (best) [21].

Coma Recovery Scale-Revised (CRS-R)
This scale includes specific ratings for several cognitive functions, including eye opening, communication skills, motor response, and environmental awareness [4].

Full Outline of UnResponsiveness (FOUR) Score
The FOUR Score is a further scale that measures four main categories: ocular response, motor response, brainstem reflex and breathing. It is used for patients with a wide range of impaired consciousness [2].

Chapter 3

Materials and Methods

3.1 Data

The data object of study of this project have been provided by "Azienda Ospedaliero-Universitaria Città della Salute e della Scienza di Torino", to the "Ontonix" firm and to the "Dipartimento di Scienze Applicate e Tecnologie of Politecnico di Torino". The dataset was composed of 54 subjects subjected to EEG in the years from 2011 to 2013. Specifically, out of a total of 54 subjects, 14 were healthy subjects who in the same years spontaneously underwent EEG, whose signals were recorded to create a reference class of normal values, while 40 were patients suffering from disorders of consciousness at different levels of severity. Pathological patiens defined as "cases" were affected by severe brain injuries, caused by trauma or vascular problems (such as internal or subarachnoidal cerebral hemorrhages, ischemic strokes), which led to a persistent coma for at least 7 days in the acute post-trauma phase, followed by a state of minimal alertness or lack of conscious response. Protocol included some exclusion criteria as age of subjects under 19 years or over 80 years, lack of sufficient clinical or anamnestic data at the time of EEG evaluation and frequent presence of epileptiform abnormalities in the EEG, in particular indicating status epilepticus. The EEG signals were also acquired in two conditions, with eyes open (with the presence in the file name of the acronym OA which stands for "occhi aperti") and with eyes closed (acronym OC which stands for "occhi chiusi"). They came in two forms, raw (located in the "interi" folder) and cut (located in the "tagliati" folder). The cut signals were previously cleaned by removing them extremely noisy or abnormal parts. All signals had the same duration of 256 seconds, sampled at 256 Hz. The first choice made concerned the dataset to use. It was chosen to work with raw time series because the cut ones suffered from strict removal of the noisy parts, which could have influenced the analysis. Although the raw signal could have included anomalous parts, adequate pre-processing has been applied to overcome this problem. As regards the acquisition condition, it was decided to include both patients with eyes closed and those with eyes open even though the recordings made with eyes open could have been influenced by any type of visual stimulus and blinking artifacts. Tracings in eyes closed mode were not available for some subjects as the consciousness of some patients was very compromised, so much so that it was difficult to ask them to keep their eyes closed. Data were organized in three different folders:

- One for Open-Eyes condition ("subject number_OA", e.g., "1_OA").
- One for Closed-Eyes condition ("subject number_OC", e.g., "1_OC").

- One for subjects with only one condition (usually Open-Eyes).

3.2 Channel selection

The recordings of the EEG signals were acquired with a non-homogeneous set of electrodes for all patients. Therefore, to ease our analysis and to ensure consistency we have eliminated the acquisitions of uncommon and non-informative channels from the files. A subgroup of 17 channels has been selected as shown in Figure 3.2. Figure 3.1 was obtained using the plot_topography.m function, showing the final position of the channels that satisfied the inclusion requirements.

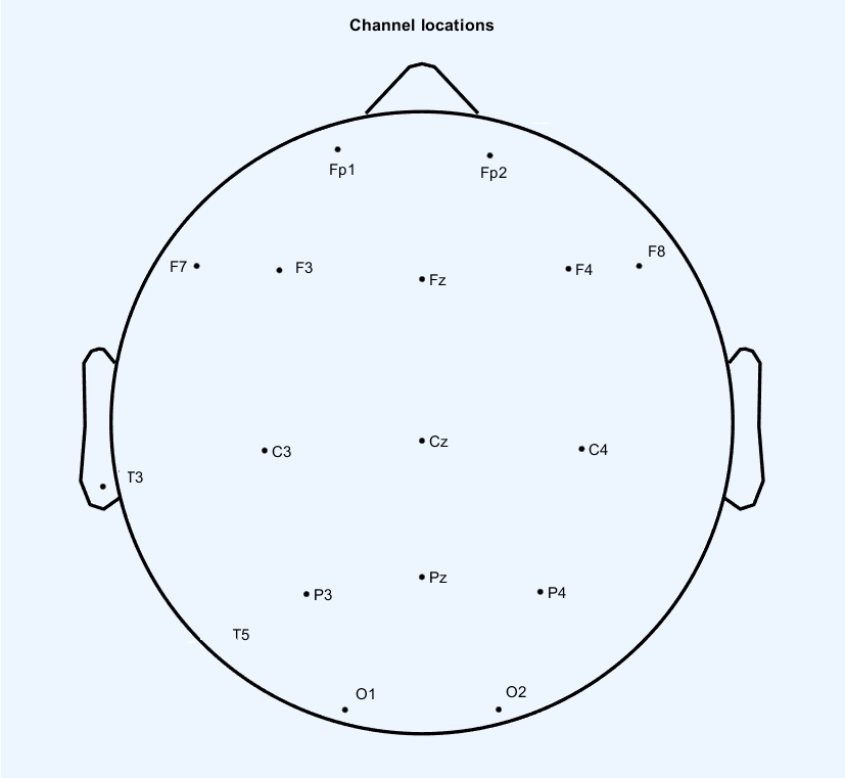

Figure 3.1: Position of the electrodes on the scalp.

Figure 3.2: Channels used for each subjects for the purpose of eliminating the non-informative and uncommon ones.

3.3 Pre-processing

3.3.1 Downsampling

The first pre-processing operation performed is the downsampling of the EEG data from the original sampling rate of 256 Hz to the final one of 16 Hz, used throughout our analysis. This operation is aimed at reducing the data volume, decreasing the storage space, allowing to make the data more manageable and obtaining a better computational efficiency with a reduction in processing times.
The effects of this operation can be observed in the following figure that show the signal of a subject and a random channel before and after the downsampling.

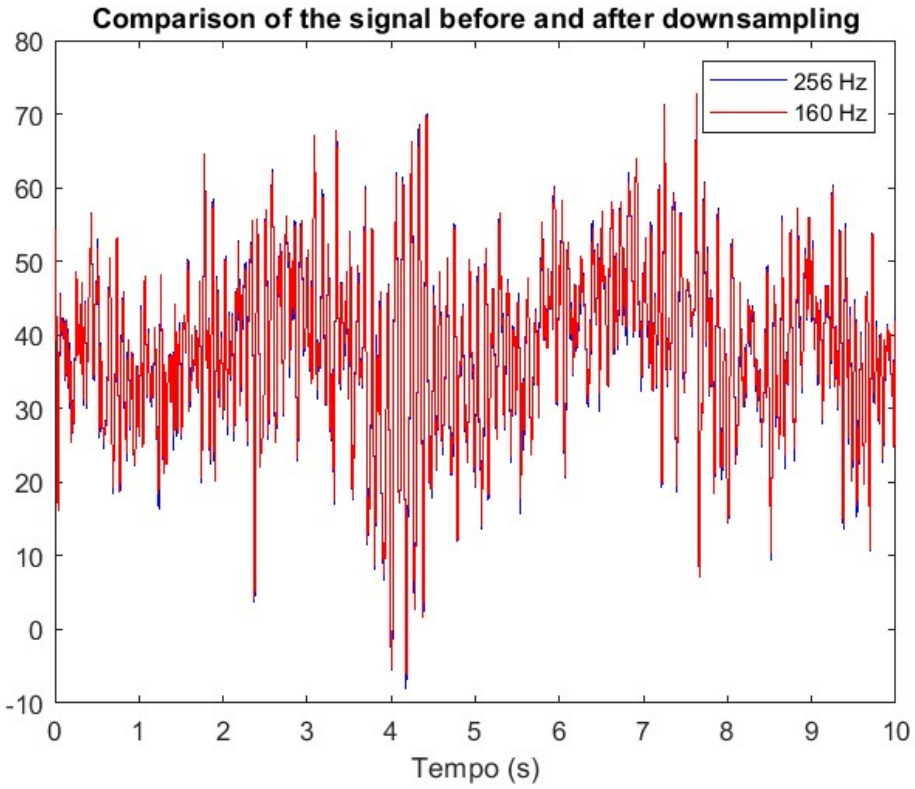

Figure 3.3: Comparison of the EEG signal of a random subject of a random channel before and after downsampling.

3.3.2 Filtering

The signals were filtered with a band-pass filter of order in the frequency range 0.5 Hz - 40 Hz using the BandPassFilter.m function. The plots of the channels of the first subject are shown below, respectively of the unfiltered signal and the post-filtering signal. The upper cut-off frequency was set to 40 Hz because the Gamma band (> 30 Hz) is considered not very informative for the type of patients we have available, as it is associated with cognitive processes of learning, memory and states of high consciousness.
The effects of the filtering can be seen in Figure 3.4.

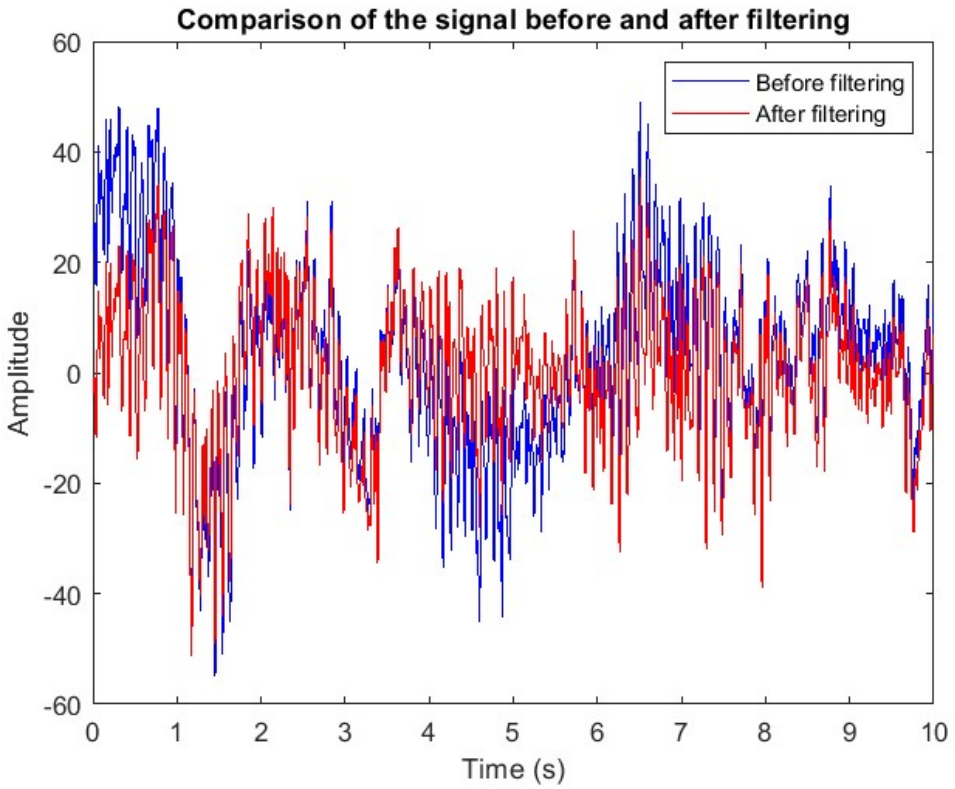

Figure 3.4: Comparison of the EEG signal of a random subject of a random channel before and after Band Pass Filtering.

3.3.3 Artefact detection

Gomez's algorithm was adopted to remove EOG and EMG artifacts. It should be noted that the EOG artefact may have already been avoided in patient acquisitions carried out in eyes closed mode, or present in a reduced manner. Even regarding EMG artifacts, a large part of them has previously been attenuated by low-pass filtering with the removal of high frequencies.
The effects of the artefacts removal can be observed in the Figure 3.5, Figure 3.6 and Figure 3.7.
The algorithm was implemented using the autobss.m function, with the following parameters:

```
    % Removal of EOG artifact
    opt_eog.wl = 200 * fs_new;
    opt_eog.ws = opt_eog.wl;
    opt_eog.bss_alg = 'sobi';
    opt_eog.bss_opt = [];
    opt_eog.crit_alg = 'eog_fd';
    opt_eog.crit_opt.wl = 0.1 * opt_eog.wl;
    opt_eog.crit_opt.ws = opt_eog.crit_opt.wl;
    opt_eog.crit_opt.method = 'sevcik_mean';

    [single_output_eog] = autobss(eeg, opt_eog);
    output_eog{subject} = single_output_eog;

    % Removal of EMG artifact
    opt_emg.wl = 3.12 * fs_new;
    opt_emg.ws = opt_emg.wl;
    opt_emg.bss_alg = 'bsscca';
    opt_emg.bss_opt = [];
    opt_emg.crit_alg = 'emg_psd';
    opt_emg.crit_opt.femg = 15;
    opt_emg.crit_opt.fs = fs_new;
    opt_emg.crit_opt.ratio = 10;

    [single_output_emg] = autobss(eeg, opt_emg);
    output_emg{subject} = single_output_emg;
```

Listing 3.1: Parameters used in Gomez's algorithm.

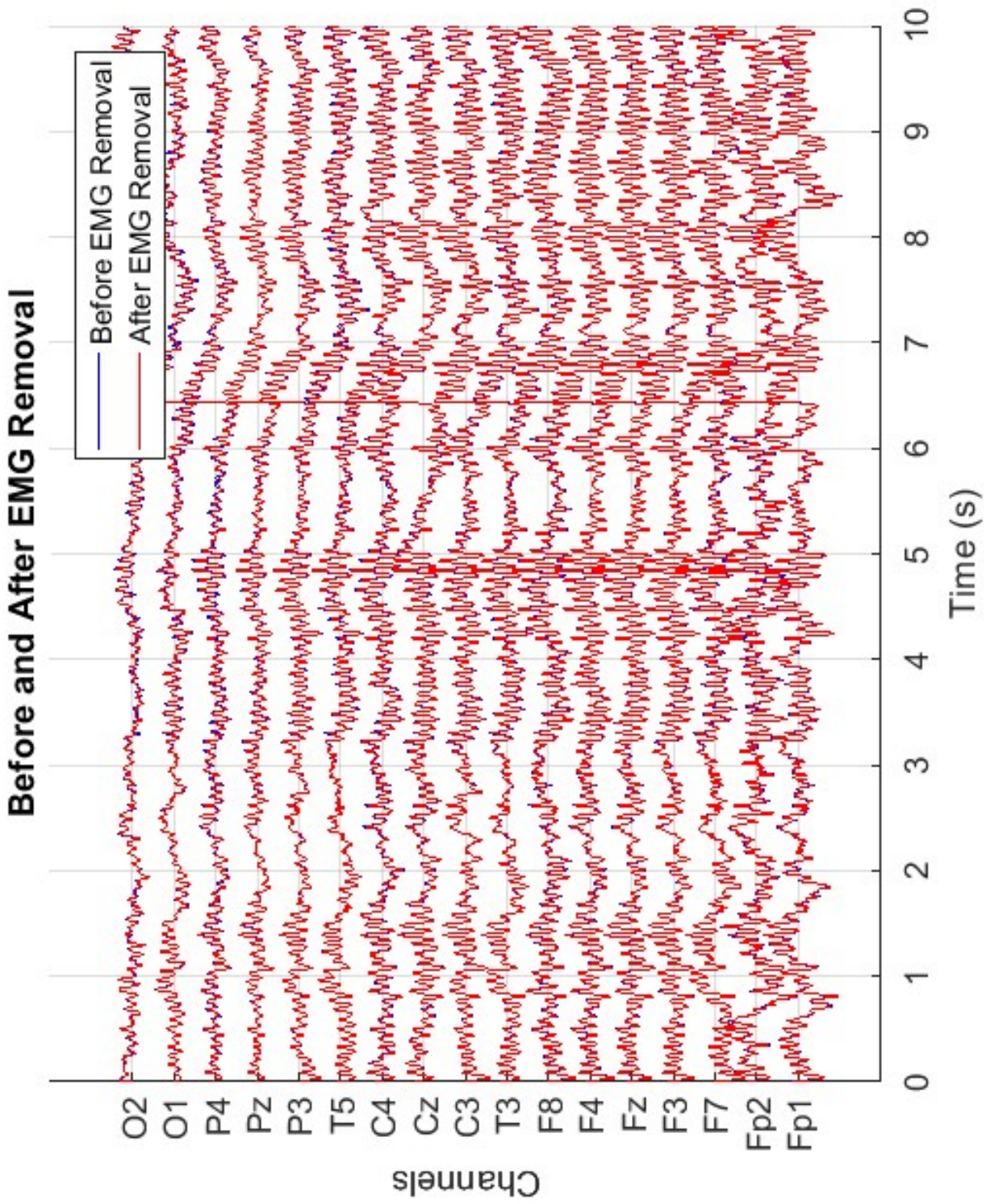

Figure 3.5: Comparison of the Raw EEG and EEG after EMG artefact removal of a random subject.

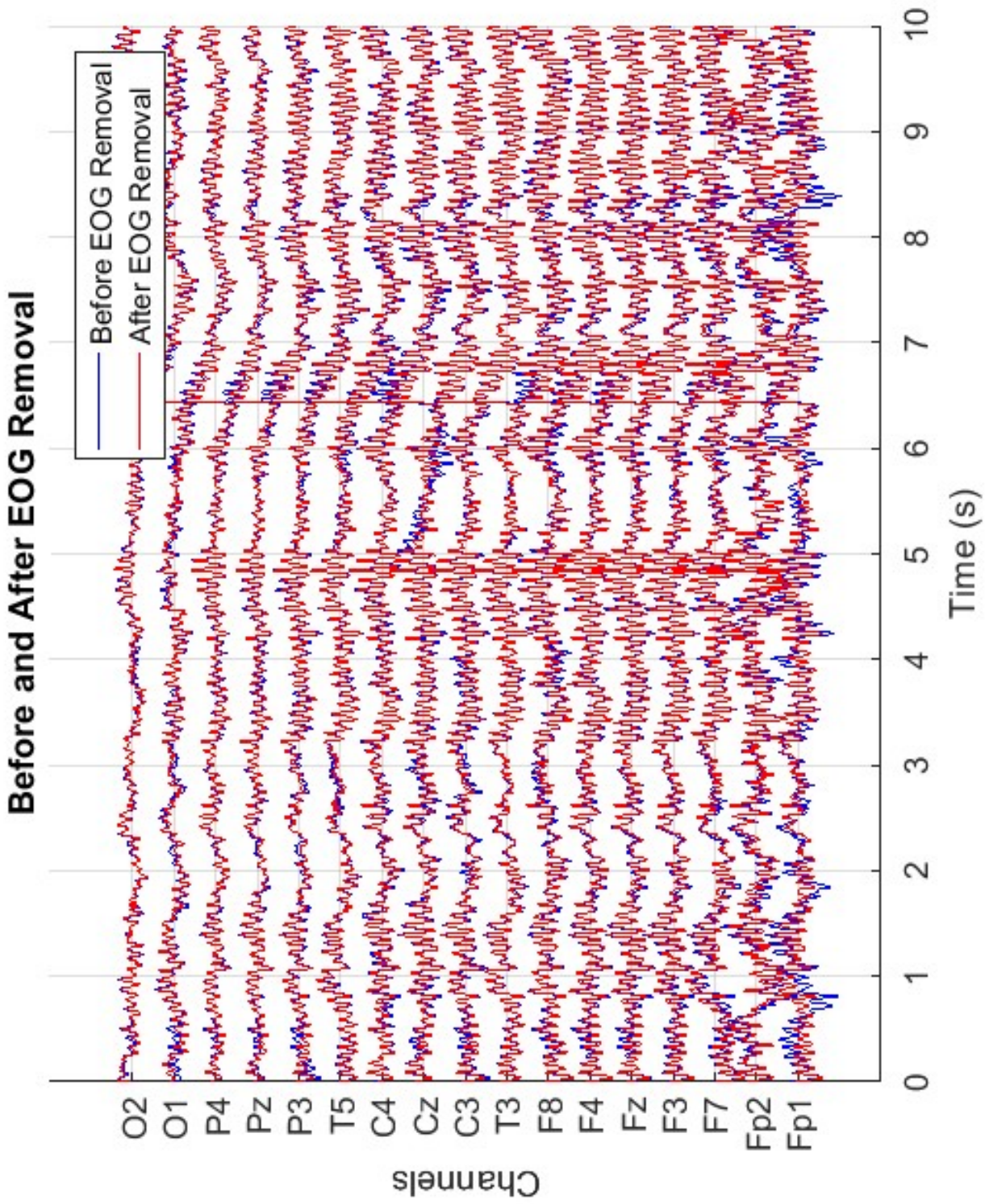

Figure 3.6: Comparison of the Raw EEG and EEG after EOG artefact removal of a random subject.

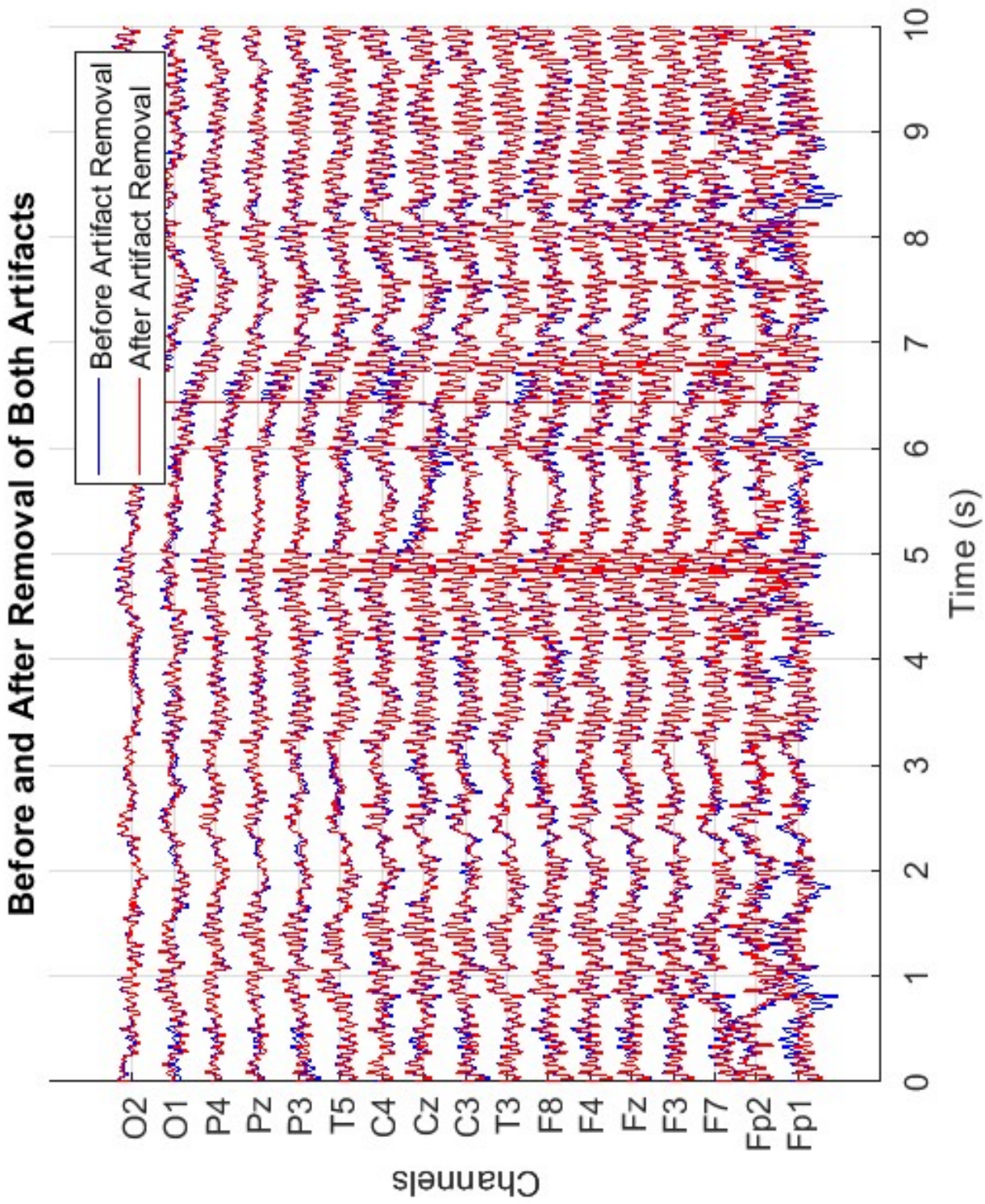

Figure 3.7: Comparison of the Raw EEG and EEG after EMG+EOG artefact removal of a random subject.

3.3.4 Epochs division

The signals were divided into 51 epochs lasting 5 seconds. This choice was adopted to facilitate the calculation of the subsequent parameters and features, thus guaranteeing the stationarity of the signal, a fundamental prerequisite for the processing of electroencephalographic signals. In themselves, in fact, EEG signals having a long duration are not stationary and their statistical properties change over time, on the other hand, on a small scale they can be considered almost stationary. The chosen time duration is the result of a compromise between the need to not have excessive amounts of data that, therefore long processing times, and the need to have good temporal and spectral resolutions.

3.3.5 Frequency bands division

Subsequently, all signals were again band-passed to isolate all frequency bands: Delta (0.5 - 3.9 Hz), Theta (4 - 7.9 Hz), Alpha (8 - 12.9 Hz) and Beta (13 - 40 Hz). Figure 3.8 depicts all frequency bands merged into the entire signal.

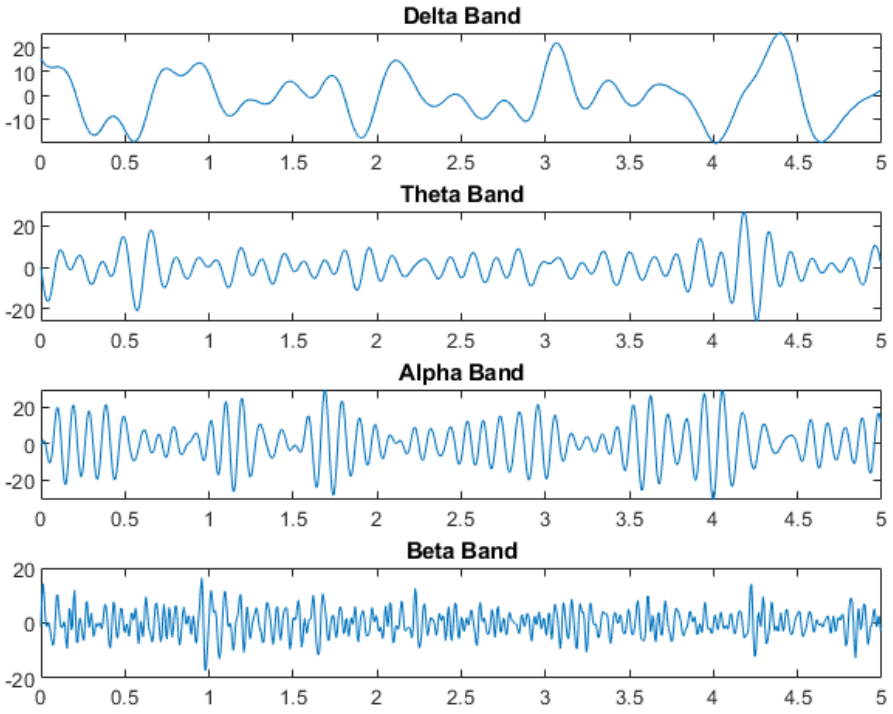

Figure 3.8: Plot of all the frequency bands for one channel of a random patient.

3.4 Features calculation

3.4.1 Time-domain features

The following statistical features in the time domain were calculated for each extracted epoch and inserted into the MATLAB file *time_domain_statistics.mat*.

Mean
The mean is the sum of all the values of the epoch divided by the total number of samples it is made up of. It gives me the measure of the central value of the epoch.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.1}$$

Variance
It gives me a quantification of the dispersion of values within the epoch with respect to the mean. Mathematically it is the average of the squares of the differences between each value and the mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{3.2}$$

Standard deviation
This is also a measure of dispersion and is represented by the square root of the variance.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{3.3}$$

Kurtosis
It is a quantification of how much the distribution of values is concentrated around the mean. In fact, high values of kurtosis indicate a distribution with high peaks and long tails.

$$K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^4 - 3 \tag{3.4}$$

Skeweness
It is a measure of the symmetry (or asymmetry) of the signal with respect to the mean. Positive values indicate an asymmetric distribution with a longer tail on the right, negative values represent the asymmetry of the values with a longer tail on the left.

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^3 \tag{3.5}$$

Range
It represents the difference between the maximum and minimum value of the signal and therefore a measure of the peak-to-peak value.

$$R = \max(x) - \min(x) \tag{3.6}$$

3.4.2 Hjorth parameters

The following parameters are useful for understanding the temporal dynamics of EEG signals without taking a transformation into the frequency domain. The Hjorth parameters have been inserted into the MATLAB file *Hjorth_Parameters.mat*

Activity
Activity is a measure of the variance of the signal over time. It represents the energy of the EEG signal and indicates the overall amplitude of the brain's electrical activity.

Mobility
It measures the rate of change of the signal, and is calculated as the square root of the ratio between the variance of the first derivative of the signal and the variance of the signal itself. It therefore describes the average frequency of the signal. It provides indications of the predominance of high or low frequency components.

$$\text{Mobility} = \frac{\sigma_{\text{derivata\_segnale}}}{\sigma_{\text{segnale}}} \tag{3.7}$$

Complexity
It measures the change in Mobility over time and is an indicator of the change in the dominant frequency of the signal. Mathematically it is calculated as the ratio of the Mobility of the first derivative of the signal to the Mobility of the signal itself. This measure reflects the amount of change in the waveforms of the EEG signal.

3.4.3 Frequency-domain features

The following parameters in the frequency domain were calculated for each extracted epoch and inserted into the MATLAB file *frequency_domain_parameters.mat*.

Average Rectified Value
ARV is a measure of the average signal magnitude, obtained by taking the absolute value of the EEG signal and averaging it.

$$\text{ARV} = \frac{1}{N}\sum_{i=1}^{N}|x_i| \tag{3.8}$$

Root Mean Square
It represents the actual magnitude of the signal, it refers to a measure of its amplitude as a whole, which takes into account not only the magnitude of the individual values of the signal, but also their variability over time. The RMS is useful for quantifying the overall energy of the EEG signal.

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2} \tag{3.9}$$

Mean Frequency
The MNF was calculated with the *fmean.m* function and represents the mean frequency of a signal in the frequency domain and indicates the dominant frequency.

$$\text{MNF} = \frac{\sum_f f \cdot P(f)}{\sum_f P(f)} \tag{3.10}$$

Median Frequency
The MDF was calculated with the *fmedian.m* function and it is the frequency below which 50% of the total spectral power of the signal is contained.

$$\text{MDF} = f \text{ such that } \int_0^f P(f')\,df' = \frac{1}{2}\int_0^\infty P(f)\,df \tag{3.11}$$

3.4.4 Bandwidth Power calculation

When we talk about power in different frequency bands, we are referring to the spectral analysis of a signal to determine how the energy or power of the signal is spread across different frequency bands. The powers of the various frequency bands were extracted from each epoch using the Welch method, which is common for estimating the power spectral density (PSD) of the signal. These features have been inserted into the MATLAB file *powers.mat*. The frequency bands analyzed are:

- **Delta (0.5 Hz - 3.9 Hz)**: Low and slow waves.Delta activity is most prominent during deep, dreamless sleep.
- **Theta (4 Hz - 7.9 Hz)**: Waves always slow, but slightly faster than the previous ones. Mostly present in states of relaxation, drowsiness (first stages of sleep), REM phase, processing of emotions and memory activities.
- **Alpha (8 Hz - 12.9 Hz)**: Fast waves associated with states of relaxation and calm without any occupation in intense mental activities.
- **Beta (13 Hz - 40 Hz)**:Very fast waves.

3.4.5 Relative powers

They indicate the proportion of the power of a certain band compared to the total power of the signal. Mathematically it is calculated as the ratio of the power of a certain band compared to the total power. It indicates how dominant that frequency interval is and therefore the relative mental state associated with it compared to the total power of the signal. The relative powers of the delta, theta, alpha and beta bands have been calculated compared to the total power of the entire frequency spectrum of each epoch. The following features have been inserted into the MATLAB file *rel_powers.mat*.

3.4.6 Frequency-domain statistcs

As performed in the time domain, also in the frequency domain statistical parameters were extracted, including the mean, median, variance, standard deviation, skewness and kurtosis of the power spectral density of each epoch for each channel and subject. These statistcs parameters have been saved in the MATLAB file *frequency_domain_statistics.mat*.

3.4.7 Frequency band ratios

Frequency band ratios are measures that compare the spectral power of one frequency band to that of another. The following band ratios have been calculated:

- Theta/Alpha Ratio:

$$\frac{P_{\text{Theta}}}{P_{\text{Alpha}}}$$

- Beta/Theta Ratio:

$$\frac{P_{\text{Beta}}}{P_{\text{Alpha}}}$$

- Alpha/Delta Ratio:

$$\frac{P_{\text{Theta}}}{P_{\text{Beta}}}$$

- (Theta + Alpha)/(Alpha + Beta) Ratio:

$$\frac{P_{\text{Theta}} + P_{\text{Alpha}}}{P_{\text{Alpha}} + P_{\text{Beta}}}$$

- (Theta + Alpha)/Beta Ratio:

$$\frac{P_{\text{Theta}} + P_{\text{Alpha}}}{P_{\text{Beta}}}$$

The calculated features were subsequently saved in the MATLAB file *Band_ratios.mat* .

3.4.8 Non-linear features

Hurst exponent
The Hurst exponent is a measure of persistence or antipersistence in a time series. A high Hurst exponent ($H > 0.5$) may indicate a time series with high coherence or predictability in brain activity, while a low value ($H < 0.5$) may suggest greater variability or chaos in the signal. This parameter was calculated using the *estimate_hurst_exponent.m* function and saved in the MATLAB file *hurst_exp.mat*.
Lyapunov exponent
By definition, it quantifies the rate at which two initially close trajectories in a state space diverge. A positive value of the Lyapunov exponent may indicate that the EEG signal is chaotic and unpredictable, suggesting high variability and potential neurological dysfunction. Conversely, a negative value may indicate greater stability and predictability of the signal, associated with lower variability and more stable cognitive states. This parameter was calculated using the *lyapunovExponent.m* MATLAB function and saved in the *lyapunov_exp.mat* file.

3.5 Entropies calculation

Entropy is the measure used to quantify the uncertainty, variability, and complexity of a system or signal. For EEG signals, being dynamic and nonlinear systems, entropy allows to obtain useful information to identify the degree of disorder and can help to analyze different brain conditions. In our analysis, the following entropy measures were calculated:

Shannon Entropy

Shannon entropy is the most common measure of entropy and represents the uncertainty associated with a random variable. It is defined for a discrete probability distribution and is calculated as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{3.12}$$

where:

- X is a random variable with n possible states;
- $p(x_i)$ is the probability of occurrence of state x_i;
- the logarithm base can be 2 (for entropy in bits) or e (for entropy in nats).

Tsallis Entropy

Tsallis entropy is another generalization of Shannon entropy and is based on a different assumption for the measurement of uncertainty. It is defined as:

$$S_q(X) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p(x_i)^q\right) \tag{3.13}$$

where:

- q is a parameter that modulates the behavior of the entropy;
- $p(x_i)$ is the probability of state x_i.

Rényi Entropy

Rènyi entropy is a generalization of Shannon entropy and is defined as:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{n} p(x_i)^\alpha\right) \tag{3.14}$$

where:

- α is a parameter that regulates the sensitivity of the entropy. When α approaches 1, Rényi entropy reduces to Shannon entropy;
- $p(x_i)$ is the probability of state x_i.

Modified Approximative Entropy
To better understand the Modified Approximative Entropy briefly describe the classical version of Approximative entropy and analyze its limitations. Given a time series, Approximate Entropy reveals the predictability (unpredictability) of the fluctuations and its complexity. It looks for repetitive patterns that can make the signal less complex and more predictable. Therefore, it allows to predict the behavior of a signal in future moments, based on what happened in previous moments of time. Defining r as the tolerance and m as the built-in dimension, ApEn represents the logarithmic difference between the probability that two sub-sequences of length m and then two of length m+1 are similar within a margin of tolerance r. Considering a time series of length N:

$$\mathbf{X}_i^m = [x(i), x(i+1), ..., x(i+m-1)] \tag{3.15}$$

where i varies from 1 to $N - m + 1$. The distance between two vectors $\mathbf{X}_i^m$ and $\mathbf{X}_j^m$ is:

$$d(\mathbf{X}_i^m, \mathbf{X}_j^m) = \max_{k=1,...,m} |x(i+k-1) - x(j+k-1)| \tag{3.16}$$

Next, for each i, the function $C_i^m(r)$ that counts the proportion of similar vectors within a threshold r is:

$$C_i^m(r) = \frac{\text{number of } \mathbf{X}_j^m \text{ such that } d(\mathbf{X}_i^m, \mathbf{X}_j^m) \leq r}{N - m + 1} \tag{3.17}$$

The function $\Phi^m(r)$ which represents the logarithmic mean of the values of $C_i^m(r)$ is:

$$\Phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln C_i^m(r) \tag{3.18}$$

For vectors of length $m + 1$, we calculate $\Phi^{m+1}(r)$ in a similar way.
Therefore, the approximate entropy is defined mathematically as:

$$\text{ApEn}(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r) \tag{3.19}$$

The main limitations and problems that this approach entails are principally the strong sensitivity to the parameters m and r. The embedding dimension must be large enough to avoid false neighbors. Low m values could lead to projections of the trajectory towards a lower dimension. False neighbors can influence the ApEn value, but they would not be related to the complexity of the system and, therefore, would be misleading. Some guidelines for the choice of parameters recommend setting:

- m set to 2 or 3;
- epoch duration greater than or equal to 10^m or 20^m samples;
- r equal to 20 % std of the signal.

High values for the embedding dimension, however, lead to a reduced probability of finding recurrences.
Therefore, values that are too small may misunderstand the structure of the series, while values that are too large may lead to uninformative measures that are very sensitive to noise.

Moreover, if r is too large, the two series may be very similar, while if the value is set to very small, too few similar vectors may be obtained.
Furthermore, ApEn is susceptible to the value of the sampling frequency. In fact, ApEn is lower if the signal is oversampled because a linearization of the signal occurs. More samples imply that each sample is closer to the nearest one, making it more predictable and less uncertain. In fact, the maximum value for ApEn is given by ln(N-m), which increases as the sampling rate and, therefore, the value of ApEn with respect to its maximum value is reduced with oversampling. The suggested sampling rate value is around the Nyquist rate, with the embedded dimension equal to the ratio (sampling rate)/(Nyquist limit).
Besides, ApEn may be sensitive to the length of the time series. For very short series, the calculation may lead to unreliable values or may not accurately reflect the complexity of the system.
Finally, the ApEn requires the time series to be stationary, i.e. its statistics do not vary over time. For non-stationary series, ApEn may not accurately capture the overall dynamics of the system.
According to the ApEn formulation, even values that are close in time are considered close, which leads to singularities that can be removed. Self-recurrences for N-m-1 are always found, shifting the result of ApEn towards lower values. This problem is overcome in two ways:

- by introducing a Theiler window that defines a certain time interval in which no recurrences can be found.

- by introducing other entropy measures, such as Sample Entropy (SampEn) or Permutation Entropy, that can overcome some limitations of ApEn because instead of relying on the logarithm operator they rely on a summation.

Modified ApEn overcomes some of its limitations[15]:

- When considering the time delay (τ) between points, a linear correlation with the sampling rate can be adopted. Delays other than 1 compensate for oversampling, avoiding self-recurrences in a given time window.

- To avoid finding too few similar segments, the tolerance (r) can be set to a specific percentage of the recurring points found in the m dimension, instead of strictly depending on the std. In some cases, r has been set to the maximum value of ApEn, and this has shown both promising and unpromising results.

- In the two embedding dimensions, [N-(m+1) τ] is the total number of embedded vectors.

- A high-pass filter is used to eliminate low-frequency trends. It has a cutoff frequency equal to the inverse of the epoch length.

- A Theiler window is used to not consider points closer than a delay as neighbors. When a recurrence is found in dimension m but not in m+1, a single recurrence is added to the correlation integral, while points that do not have similar segments in dimension m are not directly considered.

Nevertheless, the modified ApEn formulation does not guarantee better results than the original one, because more tests would have to be performed on a wider range of conditions and data. Furthermore, in certain cases, some parameter combinations led to unstable or contradictory results. However, promising results have been obtained that demonstrate some of the main advantages of using modified ApEn:

- Compensated oversampling, since the complexity is provided independently of the sampling rate or data bandwidth.
- The embedding dimension does not have to remain low, since modified ApEn provides acceptable values even if m increases, unlike the original ApEn that would reduce the number of recurrences.
- The epoch length is an important issue for the original index, since it must remain low to maintain a quasi-stationarity condition, but at the same time, the epoch cannot be too short since the ApEn would not find enough similarities. This does not happen with the modified ApEn when the correct embedding size is chosen.
- Compared to SampEn, self-occurrences are not required, so the local behavior of the signal does not affect the results. In fact, temporally shorter signals and rare events are represented as they should be.

3.6 Spatio-temporal features

Spatio-temporal features are measures that capture both spatial and temporal variations in signals, focusing on the dependencies between different brain regions. Unlike previously computed temporal features, which measure the properties of a single signal over time, spatio-temporal features take into account the interactions between multiple signals recorded from different brain regions, thus allowing for insights into the functional connectivity of the brain. Each feature was calculated for each channel pair and organized into 4D matrices with dimensions 95 x 51 x 17 x 17. (num_subjects x num_epochs x channels x channels). The following connectivity features were calculated:

- Cross-correlation.
- Coherence.
- Lagged-Coherence.
- Mutual Information.
- Covariance.
- Pearson's Correlation Coefficient.
- Phase Locking Value.

3.6.1 Cross-correlation

It measures the degree of similarity between two time series as a function of a time delay (lag) applied to one of the two signals. High values can mean good connectivity between two different areas depending on the channel pair on which it was calculated. Low values indicate poor synchronization.It can be expressed as:

$$R_{xy}(k) = \sum_{n} x(n) \cdot y(n+k) \tag{3.20}$$

where:

- $R_{xy}(k)$ is the cross-correlation between signals x and y at lag k;
- $x(n)$ and $y(n)$ are the two signals being analyzed;
- k is the lag or time shift.

3.6.2 Coherence

Similar to correlation but measures the synchronization in terms of frequency between acquired brain electrical activity in different brain regions. In patients with disorders of consciousness, a reduction in coherence in specific frequency bands may indicate dysfunctions in communication between different brain regions. For example, low values in the alpha band may be related to reduced cognitive abilities or alertness. Coherence

may provide information on the functional connection between different brain areas. It can be expressed mathematically as:

$$C_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f) \cdot P_{yy}(f)} \tag{3.21}$$

where:

- $C_{xy}(f)$ is the coherence between signals x and y at frequency f;
- $P_{xy}(f)$ is the cross-spectral density between x and y;
- $P_{xx}(f)$ and $P_{yy}(f)$ are the power spectral densities of signals x and y, respectively.

3.6.3 Lagged-coherence

It measures the coherence between two EEG signals taking into account a time delay, excluding the influence of zero-delay components (concurrent components). It is a measure that attempts to assess functional connectivity that is not due to volume effects or instantaneous conduction. It is defined mathematically as:

$$C_{\text{lagged}}(f) = \frac{|P_{xy,\text{lagged}}(f)|^2}{P_{xx,\text{lagged}}(f) \cdot P_{yy,\text{lagged}}(f)} \tag{3.22}$$

where:

- $C_{\text{lagged}}(f)$ is the Lagged Coherence between signals x and y at frequency f;
- $P_{xy,\text{lagged}}(f)$ is the cross-spectral density between x and y, calculated considering only the lagged components;
- $P_{xx,\text{lagged}}(f)$ and $P_{yy,\text{lagged}}(f)$ are the power spectral densities of signals x and y, respectively, after removing simultaneous components.

3.6.4 Mutual Information

MI It measures the statistical dependence and the shared information between two variables, quantifying the amount of information present on one variable while observing the other. It measures the degree of connectivity even in the absence of linear relationships and evaluates how much the information contained in an EEG signal can predict or explain the behavior of another EEG signal. It is expressed by the following formula:

$$MI(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.23}$$

where:

- $H(X)$ and $H(Y)$ are the entropies of signals X and Y;
- $H(X,Y)$ is the joint entropy of X and Y.

3.6.5 Covariance

Covariance measures the degree to which two variables vary together and in particular the linear relationship between the two signals. It is a more general measure than correlation. It is defined mathematically as:

$$\text{Cov}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \tag{3.24}$$

where:

- N is the number of samples;
- x_i and y_i are the values of signals x and y at sample i;
- μ_x and μ_y are the means of signals x and y.

3.6.6 Pearson's Correlation Coefficient

Often calculated in conjunction with covariance, the Pearson correlation coefficient measures the strength and direction of the linear relationship between two variables. It is normalized between -1 and 1, with 1 indicating a perfect positive correlation, -1 a perfect negative correlation, and 0 no correlation. It can be expressed mathematically as:

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y} \tag{3.25}$$

where:

- ρ_{xy} is the Pearson correlation coefficient between X and Y.
- $\text{Cov}(X, Y)$ is the covariance between X and Y.
- σ_x and σ_y are the standard deviations of signals X and Y.

3.6.7 Phase-Locking Value

PLV measures the phase synchronization between two signals. A reduction in PLV may indicate a disturbance in neural synchronization, common in conditions such as coma or dementia, where functional connectivity is often compromised. It is expressed by the following formula:

$$PLV = \left| \frac{1}{N} \sum_{n=1}^{N} e^{j\Delta\phi(n)} \right| \tag{3.26}$$

where:

- PLV is the Phase Locking Value, a measure of phase synchronization between signals x and y;
- N is the number of samples;
- $\Delta\phi(n)$ is the phase difference between signals x and y at sample n.

3.7 Features organization

The approach involves an organization of the features through the calculation of summary statistical indicators, and therefore it was decided not to work on the raw values of the features. In particular, after calculating each feature for each epoch of the signal, and each connectivity matrix, it was decided to calculate statistical indicators of the first three orders such as mean, variance and skewness on the epochs of each channel and each subject. Therefore, the matrix of the features, which will be the input of the clustering and classification algorithms that we will test, will have a size equal to *num_subject* x *num_features* with *num_subject* equal to the number of subjects included in the analysis scenario and *num_features* the number of features multiplied by the number of channels and by the number of summary statistical indicators used.

3.8 Clustering algorithms

Clustering algorithms were applied to the extracted features to group the data based on the similarities between the different data without using the predefined labels, provided by the medical staff who performed the data acquisitions and made them available. This unsupervised learning technique is essential to find hidden patterns or natural groupings in the data, which may not be evident through visual inspection or traditional analysis methods. Through data clustering, the study aims to discover groups of patients with similar EEG characteristics, potentially corresponding to different states of consciousness with different levels of severity. The clustering process was accompanied by feature selection to identify the most relevant features to distinguish between the clusters. Feature selection is crucial because it reduces the dimensionality of the data, improving the performance of the clustering algorithm and providing a clearer interpretation of the underlying patterns.

3.8.1 KMeans

The unsupervised clustering process can be described through the following steps:

1. Random assignment of samples to k clusters (hyperparameter initially defined).

2. Calculation of the distance between each sample and all class centroids (midpoints of the clusters). Initially the centroids are initialized randomly and the number is consistent with the set k value. Different distance measures can be used, from pi

3. Assignment of each element to the cluster with the closest centroid.

4. Recomputation of centroids based on the new assignment of elements to clusters.

5. Iteration of the assignment phase and recomputation until the algorithm converges, i.e. no new assignment.

A common method for determining the value of K is to use the elbow plot, which graphs the trend of the total squared error as the hyperparameter K varies [12].

3.8.2 Agglomerative Clustering

Hierarchical clustering allows you to create a hierarchical structure between the data that can be used to subsequently define the clusters [1]. The process can be outlined through the following steps:

1. Initially, each data point represents a cluster in itself.
2. Calculate the distances for all pairs of data points available.
3. Merge the two data points with the smallest distance or the most similar.
4. Repeat the previous steps with merger of the two data points with the smallest distance until there is only one cluster that merges all the points.
5. Representation of the clusters through the dendogram in which the x-axis represents the data points and the y-axis the distance.

Among the advantages of this algorithm is the non-definition of an initial setting of hyperparameters such as the number of clusters. On the other hand, it is computationally expensive for large datasets and strong sensitivity to the similarity metric used.

3.8.3 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm, often used to identify clusters in a dataset that may contain noise (anomalous data or outliers). It requires the setting of the following parameters: Main parameters:

- **Eps**: Search radius, i.e. the maximum length of the radius of the circle that is drawn starting from a point and that determines the points close to it. In fact, a point that falls inside the circle of radius eps is defined as 'close' to the first point.
- **MinPts**: Minimum number of points required to form a cluster. If a point has at least MinPts neighbors within a radius of Eps, it becomes the "center point" of a cluster.

The following defines different 'types of points':

- **Core points**: Points that have at least MinPts neighbors within the radius of Eps.
- **Border points (Non-Core-Points)**: Points that are not themselves core points, but are within the radius Eps of a core point.
- **Noise points**: Points that neither belong to a cluster nor are close to a core point. These points are considered as noise or outliers.

For each point in the dataset, the algorithm checks whether it is a core point. If it is, a new cluster is started that includes all reachable points (i.e. all points within the distance Eps that are themselves core or boundary points).This process continues until all core and boundary points are assigned to a cluster. Points that cannot be included in any cluster are labeled as noise. The algorithm automatically identifies the number of clusters, without the need to set a default value. It is able to handle those points

that constitute noise and is able to handle very large datasets and clusters of arbitrary size. On the other hand, it have an high sensitivity to the choice of hyperparameters that determine its effectiveness and difficulty in managing clusters with different point density [10].

3.9 Features selection method

Feature selection is a machine learning and data analysis technique useful to reduce the dimensionality of the dataset, when a large number of features are available and many of them could be irrelevant. The additional advantages of this technique are:

- improvement of the model performance
- reduction of the algorithm training time
- reduction of the risk of overfitting
- simplification of the model by eliminating redundant and superfluous features.

3.9.1 SelectKBest

SelectKBest is a feature selection technique that is part of the data pre-processing phase in the field of machine learning. It evaluates the importance of the features and selects the K most relevant features based on a scoring function [6]. The main scoring functions used are:

- **Chi-squared** for categorical features.
- **F-statistics (ANOVA)** for numeric features.
- **Mutual information** measures the mutual dependence between two variables.

Therefore, the parameters to set are the scoring function*score_func* and K, the maximum number of features to select. The most common *score_func* are:

- ***f_regression*** used for linear regression problems.
- ***mutual_info_regression*** used for regression problems and calculates the mutual information between two random variables.
- ***f_classif*** used for classification problems.
- ***mutual_info_classif*** used for classification problems and calculates mutual information.
- ***chi2*** used for classification problems and calculates chi-square statistics.
- ***SelectPercentile*** used to select the highest X

3.10 Classification Methods

To classify EEG signals between healthy subjects and subjects with different levels of disorders of consciousness, some machine learning algorithms have been implemented to find the most performing and effective one. Both supervised and unsupervised algorithms have been tested to compare the different performances. Through our dataset, the algorithm has been trained to recognize and correctly classify in a second moment the data never seen by the model. For our purpose, the classification is not binary, between healthy and sick subjects but multiclass, because in addition to the category 'healthy subjects', labeled with the code '0', we have 4 different classes of subjects with disorders of consciousness with different levels of severity (labeled with a code from 1 to 4).

3.10.1 Multinomial Logistic Regression

The Multinomial Logistic Regression method is an extension of the Logistic Regression algorithm (native for binary classifications), when the goal is to assign data to more than two classes [13]. This algorithm uses an approach based on a linear combination of the features as a discriminant function of each class, then the softmax function transforms the generated scores into probabilities of belonging to a certain class and the log-likelihood function as the objective function to be maximized.

Definition of variables
Let $X = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ be the input vector with n features.
Let $Y \in \{1, 2, \ldots, K\}$ represent the dependent variable (class), where there are K possible classes.
Let $\theta_k = (\theta_{k0}, \theta_{k1}, \ldots, \theta_{kn}) \in \mathbb{R}^{n+1}$ represent the model's coefficients for class k, including the intercept θ_{k0}.

Linear Discriminant Function
Instead of modeling the probability directly for each class, a linear discriminant function is used for each class k:

$$\eta_k(X) = \theta_{k0} + \theta_{k1}x_1 + \theta_{k2}x_2 + \cdots + \theta_{kn}x_n = \theta_k^T X$$

where $\eta_k(X)$ is the linear score for class k.

Class probabilities using the Softmax Function
The probability of belonging to each class is modeled using the softmax function. The probability that an observation X belongs to class k is given by:

$$P(Y = k \mid X) = \frac{e^{\eta_k(X)}}{\sum_{j=1}^{K} e^{\eta_j(X)}}$$

where:

- $\eta_k(X)$ is the linear value (logit) associated with class k for observation X. It is often calculated as a linear combination of the characteristics of X
- $e^{\eta_k(X)}$ is the exponential of that score.
- The denominator normalizes the probabilities to sum to 1.

This type of strategy represented by the flowchart allows us to better understand the decision-making process, allowing us to understand the reason for the decisions taken.

3.10.2 Random Forest

Random Forest is a supervised learning algorithm that consists of a set of decision trees. It is an improved extension of the previous algorithm, which tries to solve some of its main limitations, such as overfitting and instability. The key idea of Random Forest is to build several independent decision trees during the training process and combine their results to obtain a more robust and accurate prediction. This is done through two fundamental techniques: bagging and feature randomization [5]. In bagging, a random subset of the original dataset is created (bootstrap method). This means that each tree is trained on a different sample of the data, allowing for repetitions. This helps to reduce the overall variance and prevent overfitting. The top-rated class becomes the final prediction. With feature randomization, in addition to sampling the data, Random Forest also randomly selects features for each tree. Instead of considering all the features available at each node, the algorithm chooses only a random subset of them.

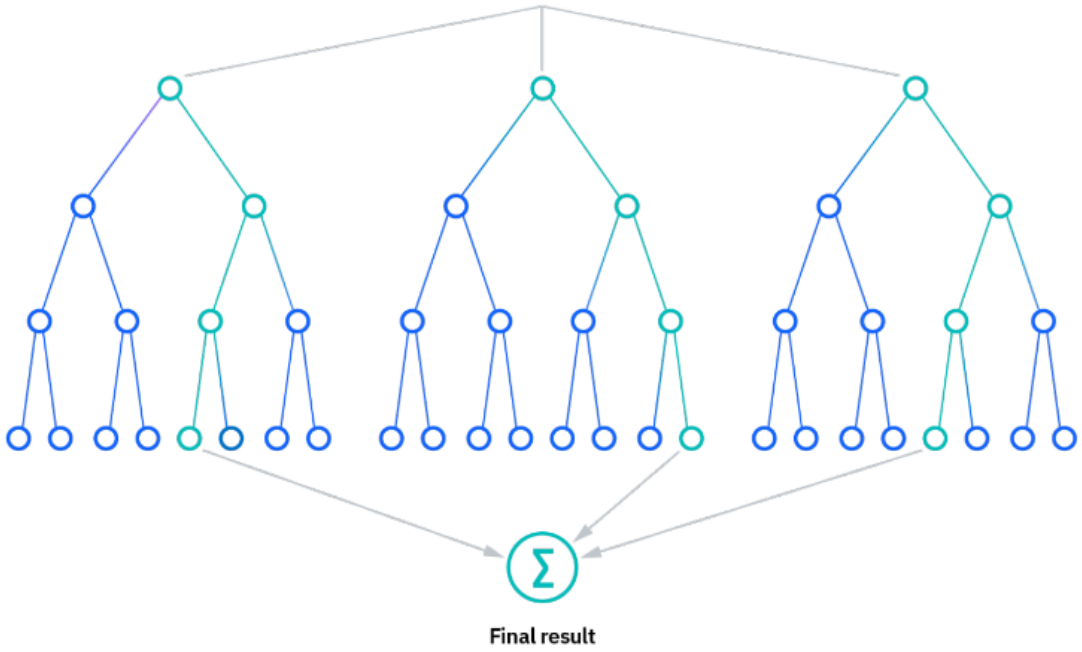

Figure 3.9: Example of structure of the Random Forest.

3.10.3 Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression problems. Its main goal is to find an optimal hyperplane that separates the data into different classes, with the goal of maximizing the margin between the data points closest to the separation boundary, known as support vectors. Having a larger margin generally results in better generalization of the model [22].

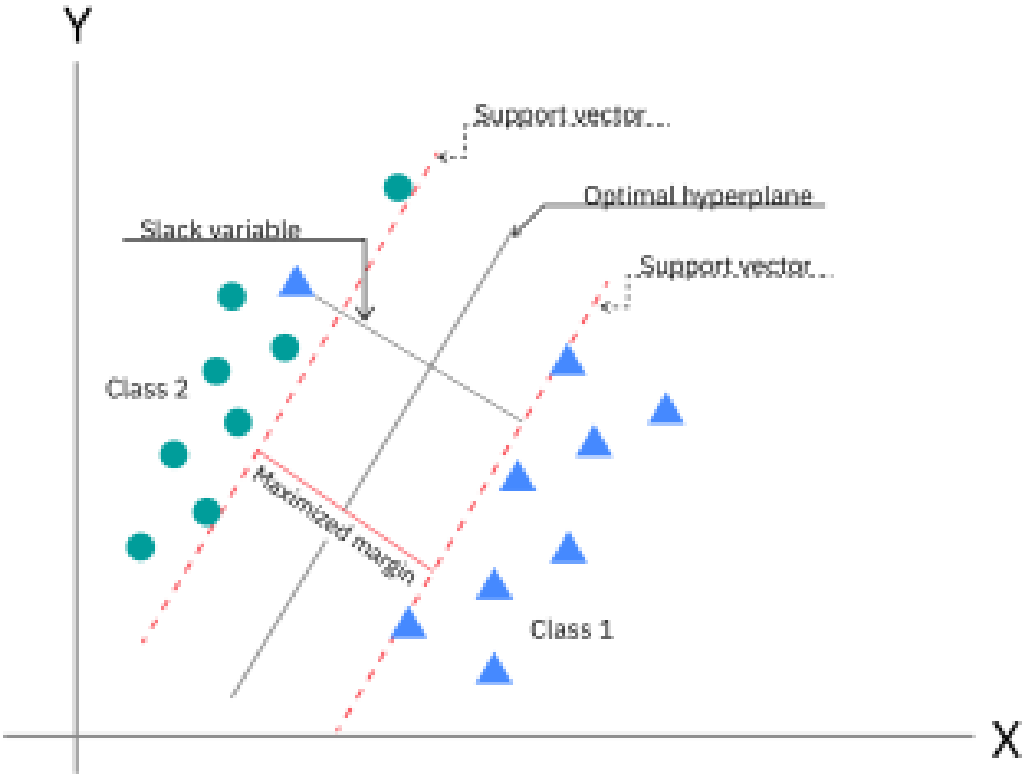

Figure 3.10: Illustration of the best hyperplane that separates two classes in a 2D chart.

If the classes are linearly separable, the goal of the SVM is to find the hyperplane that maximizes the distance (margin) between the classes. The equation of the hyperplane in an n-dimensional space is given by:

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{3.27}$$

where:

- $\mathbf{w}$ is the vector of coefficients (weights) that define the hyperplane.
- b is the bias term (intercept),
- $\mathbf{x}$ is the feature vector (or a point in the feature space).

When the data is not linearly separable, the SVM introduces a penalty term to allow for some classification error, trying to achieve a trade-off between maximum margin and classification errors.

Chapter 4

Procedure

The data were loaded from the respective MATLAB files, containing both the features in numeric array format and the associated labels. Subsequently, the possible presence of missing values (Nan) was checked and an imputation strategy was applied, if necessary, based on the replacement of missing values with the mean of the columns using the *SimpleImputer* tool of Scikit-learn. To avoid that some features had a greater weight than others, for example due to different scales, a normalization of the features was applied using the *Normalizer* tool, which reduces the feature values on a common scale based on the L2 norm. Due to the high number of extracted features, the feature selection technique called SelectKBest based on the mutual information between the features and the labels was applied. This technique selects the most informative features by eliminating influential ones. It was experimented with different values of the parameter k, representing the number of features to extract. The parameter k was set to cover a range between 4 and 30. After selecting the best features, the Agglomerative Clustering, KMeans and DBSCAN algorithms were applied.

Agglomerative Clustering and KMeans
For the first two algorithms, the parameter indicating the number of clusters (*n_clusters*) is set. Knowing the number of classes in which the data is divided, the parameter was set to 5, for the 'all', 'OA' and 'OC' scenarios. While, for the 'Only_OA' scenario, characterized by a much lower number of data samples, with the possibility of not having a sample available for each class within the dataset, *n_clusters* was set to a range between 2 and 5.

DBSCAN
Different configurations of the *eps* parameters (0.1, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0) and *min_samples* (2, 3, 4, 5, 8, 10, 12, 15) were explored, looking for the optimal combination that maximized the separation of the clusters.

Three metrics were calculated to evaluate the quality of the clustering:

- **Silhouette Score**: measures how well the points in a cluster are similar to each other compared to the points of other clusters. Higher values indicate a better separation between the clusters.

- **Davies-Bouldin Index (DBI)**: measures the compactness and separation of the clusters. A small value indicates a better configuration.

- **Calinski-Harabasz Index**: evaluates the density of clusters with respect to their separation. Higher values indicate better performance.

The feature selection cycle was repeated for all combinations of k and possibly *n_clusters*, for the first two algorithms, and of *eps* and *min_samples* for DBSCAN, and finally the best combination that maximized the Silhouette Score value was extracted. After the extraction of the best combination of parameters, each algorithm was retrained with the set of features extracted from SelectKBest.

With the features extracted from the clustering algorithm that provided the best values of the performance indices, different classification algorithms belonging to the category of supervised learning methods were applied, with the aim of predicting the class of membership of the samples constituting the dataset. The tested algorithms are Random Forest, Support Vector Machine and Logistic Regression. Initially, a pre-processing phase of the features extracted in the previous phase is planned.
Through the *PowerTransformer* method, the Yeo-Johnson transformation was applied with the aim of reducing the variance and eliminating any non-uniform distributions. This phase improves the model's ability to generalize on unknown data. Subsequently, *MinMaxScaler* was applied with the aim of normalizing the data by reducing them to a range between 0 and 1. Through a preliminary analysis of the dataset, an imbalance of the classes was noted, as some are represented by a much smaller number of samples than the others. It was decided to use the *SMOTE* (Synthetic Minority Over-sampling Technique) technique with the aim of generating synthetic samples for the minority classes through an approach based on k-nearest neighbors, although this technique is mostly used in image analysis and processing. In the code, *k_neighbors=2* was set. This parameter indicates the number of closest samples to be considered to generate the synthetic samples of the minority class. The resulting balanced dataset is then divided into train and test sets with the *train_test_split* function, foreseeing 80% of the data for training the model and 20% for evaluation. To find the best configuration of the classifier parameters, the Grid Search technique with 5-fold cross-validation (CV) (KFold) was used.
Parameters considered for tuning the Random Forest algorithm include:

- ***n_estimators***: number of trees in the forest.
- ***max_depth***: maximum depth of trees.
- ***min_samples_split***: minimum number of samples required to split an internal node.
- ***min_samples_leaf***: minimum number of samples required to be a leaf.
- ***bootstrap***: whether or not to bootstrap the data.

The parameters used for tuning the SVM algorithm are:

- ***C***: The regularization parameter C controls the trade-off, that is, the relationship between the classification accuracy on the training set and the ability of the model to generalize. The values tested are: 0.1, 1, 10, 100.
- ***gamma***: controls the influence of a single training sample. A high value of gamma indicates that the model fits the individual points very well, while a low value means that the influence of a single point is further away. The values tested are *'scale'*

which sets the gamma parameter as the inverse of the product of the number of features and the variance of the dataset features; and *'auto'* which sets it as the inverse of the number of features.

- ***kernel***: Describes the function used for data transformations in the feature space. The values tested are: *'rbf'* to set nonlinear kernels and *'linear'* for linear kernels.

Whereas, the tuning parameters for the Logistic Regression algorithm are:

- ***C***: controls the regularization. The values tested are: 0.1, 1, 10, 100.
- ***penalty***: L2 regularization (ridge) is used, penalizing large weights.
- ***solver***: Specifies the optimization algorithm; *'lbfgs'* is suitable for multiclass problems, while *'liblinear'* is often used for binary classification problems.

The *GridSearchCV* function has the function of evaluating each combination of parameters based on accuracy, selecting the one with the highest value in the cross-validation phase. As for After identifying the best parameters, the optimized model is trained on the entire training set. Predictions on the test data are run using the *predict* function to obtain the predicted classes and with *predict_proba* to obtain the predicted probabilities for each class.

Subsequently, the following evaluation metrics are computed:

- ***Accuracy***: the percentage of correct predictions on the test set.
- ***Multiclass AUC***: the area under the ROC curve for multiclass problems, using the *one-vs-one* (OVO) approach with a weighted average for each class. The AUC provides a measure of the model's ability to distinguish between different classes.

A classification report is also provided that includes the precision, recall, and F1 score measures computed for each class. Finally, to visually evaluate the model performance, the ROC curve is plotted for each class. The ROC curve shows the relationship between the false positive rate (FPR) and the true positive rate (TPR) for each class. The area under the curve (AUC) is also calculated for each curve, and a graph shows the different curves, allowing performance comparisons between classes.

Chapter 5

Results

5.1 Clustering Results

For the three analysis contexts, three clustering methods were evaluated: Agglomerative Clustering, KMeans, and DBSCAN, calculating metrics such as the Silhouette Score, the Davies-Bouldin Index, and the Calinski-Harabasz Index as reported in the following tables.

5.1.1 Clustering performances - Scenario 'ALL'

The following table shows the values of the different performance evaluation parameters of the clustering methods used.

Method	**Silhouette Score**	**Davies-Bouldin Index**	**Calinski-Harabasz Index**
Agglomerative Clustering	0.9278	0.0	101634.5738
KMeans	0.9467	0.1242	9320.7024
DBSCAN	0.9464	0.1734	508.3439

Table 5.1: Values of performance indicators for the three clustering types referred to the 'ALL' scenario.

It is possible to notice how KMeans and DBSCAN have provided almost similar results in terms of Silhouette Score. Regarding the use of these two algorithms that have provided the best results, we can highlight some advantages and disadvantages. The KMeans algorithm requires the setting of the n_clusters parameter, which could be considered a disadvantage if you do not know the clusters present in the dataset, but in our case, you know the number of existing classes. The DBSCAN algorithm is very useful, however, in cases where this information is not available. Furthermore, KMeans is limited in the search for spherical clusters. The DBSCAN algorithm is very effective in managing noise and outliers, but may have difficulties in cases where the density within the clusters varies.

5.1.2 Cluster plots - Scenario 'ALL'

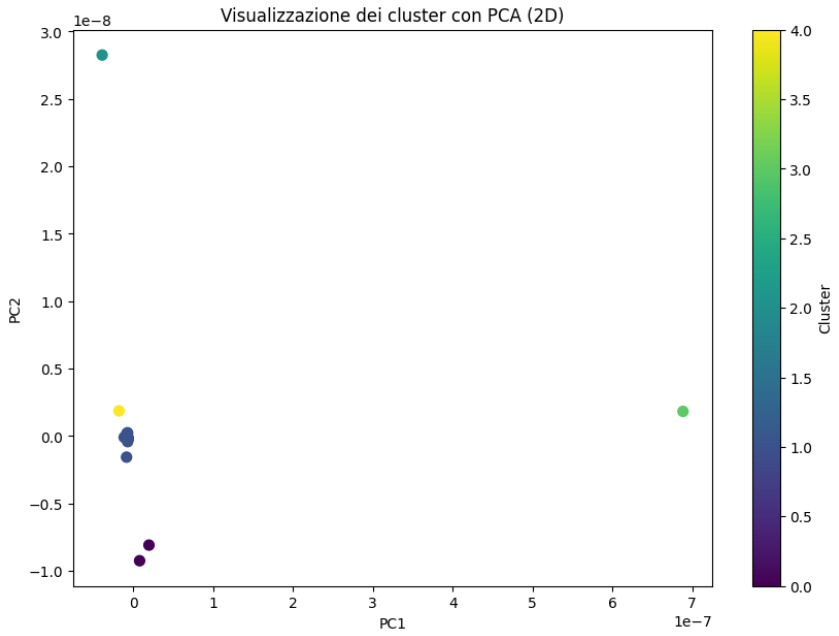

Figure 5.1: Cluster representation with Agglomerative Clustering method.

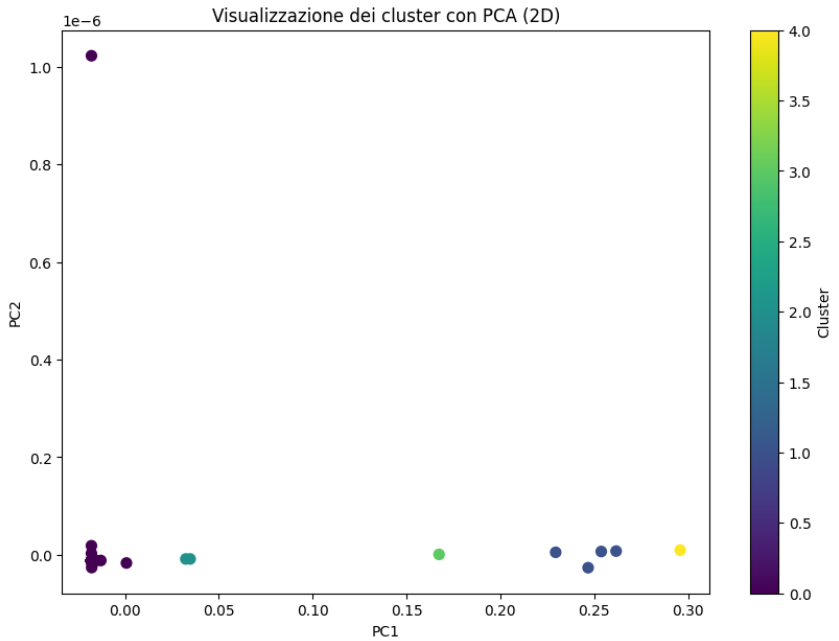

Figure 5.2: Cluster representation with KMeans method.

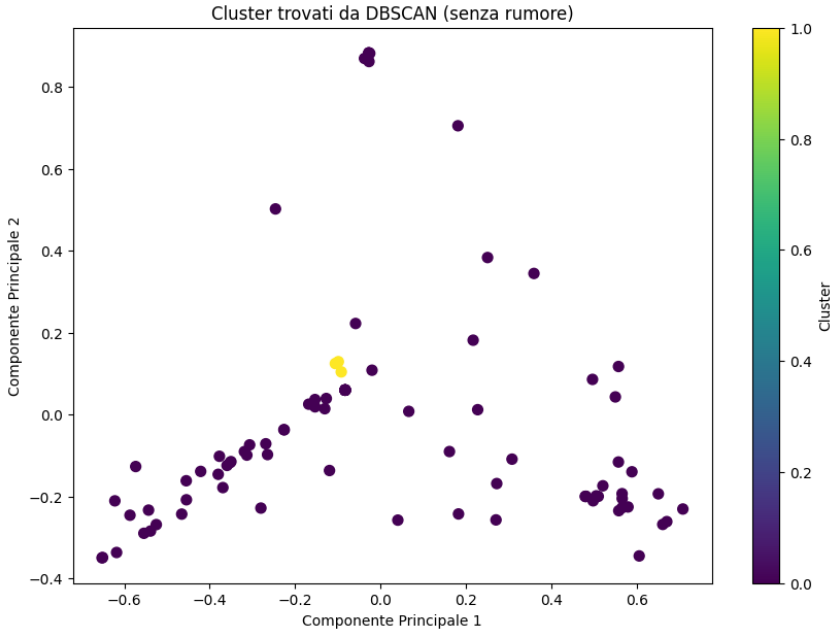

Figure 5.3: Cluster representation with DBSCAN method.

5.1.3 Clustering performances - Scenario 'OA'

Method	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Agglomerative Clustering	0.8902	0.1152	30805.50079
KMeans	0.8929	0.1130	5398.1257
DBSCAN	0.8560	0.2037	267.2200

Table 5.2: Values of performance indicators for the three clustering types referred to the 'OA' scenario.

In this scenario, despite all three methods maintaining very high Silhouette Score values, KMeans algorithm again proved to be the best. While DBSCAN had difficulty in ensuring good results in terms of separability, dividing the dataset into only 2 clusters again, and in terms of density, with a higher Davies-Bouldin Index than the other two.

5.1.4 Cluster plots - Scenario 'OA'

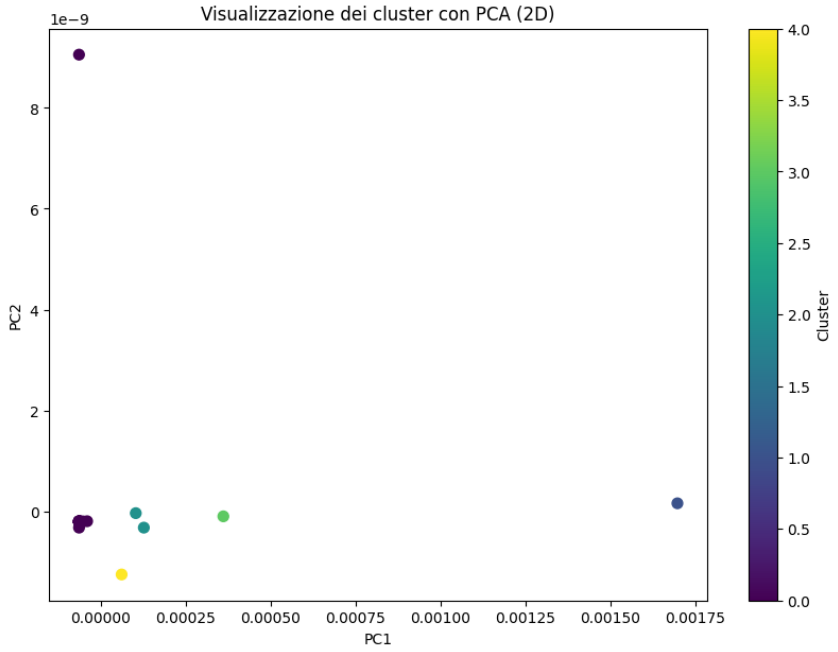

Figure 5.4: Cluster representation with Agglomerative Clustering method.

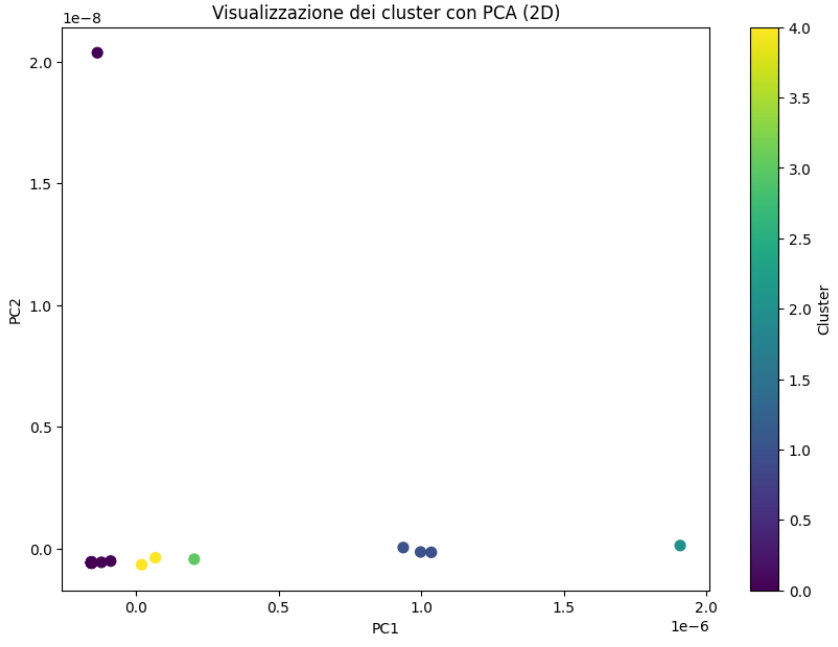

Figure 5.5: Cluster representation with KMeans method.

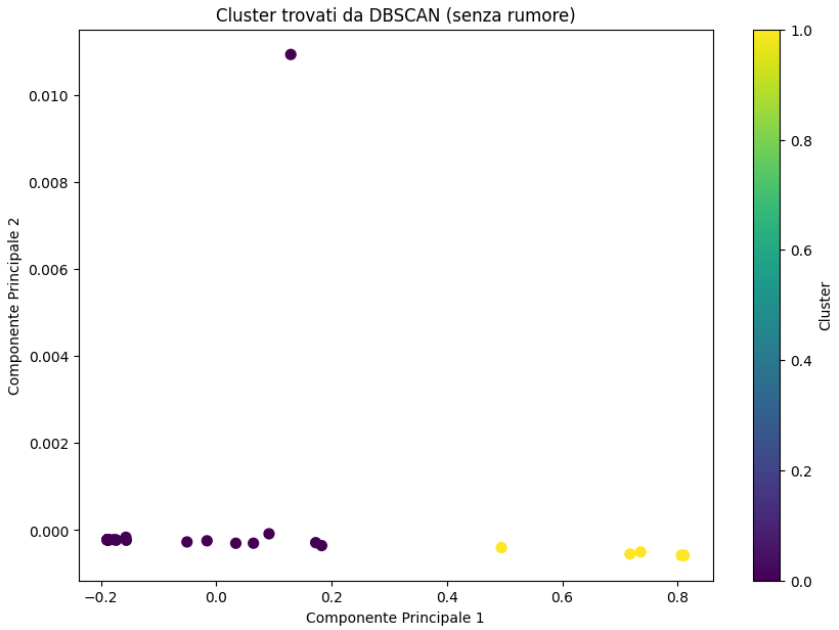

Figure 5.6: Cluster representation with DBSCAN method.

5.1.5 Clustering performances - Scenario 'OC'

Method	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Agglomerative Clustering	0.8706	0.1806	102715.2469
KMeans	0.8676	0.1103	980315.5051
DBSCAN	0.5960	0.2909	16.5907

Table 5.3: Values of performance indicators for the three clustering types referred to the 'OC' scenario.

Finally, in the last scenario, it is possible to observe a greater variability in the results with Agglomerative Clustering and KMeans that produced similar values in terms of Silhouette Score (0.8706 and 0.8676), but KMeans showed the best Davies-Bouldin Index and Calinski-Harabasz Index, making it more effective in managing data separation compared to DBSCAN, which had difficulty in distinguishing clusters, obtaining poor values for all three parameters.

5.1.6 Cluster plots - Scenario 'OC'

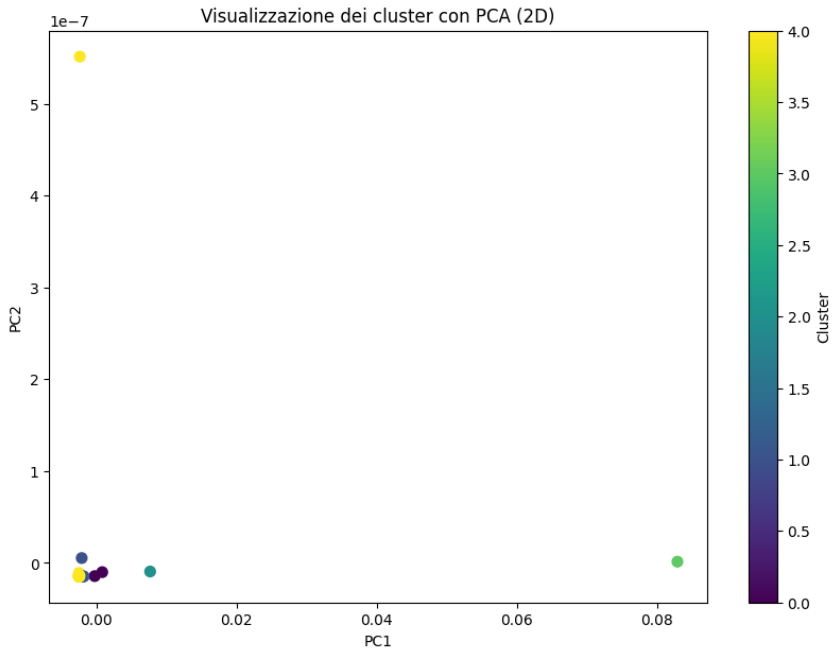

Figure 5.7: Cluster representation with Agglomerative Clustering method.

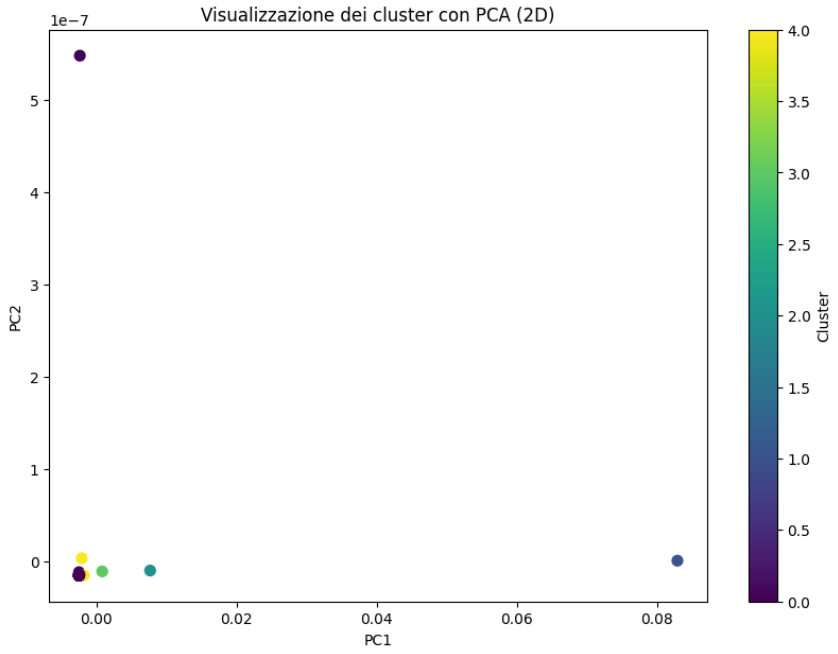

Figure 5.8: Cluster representation with KMeans method.

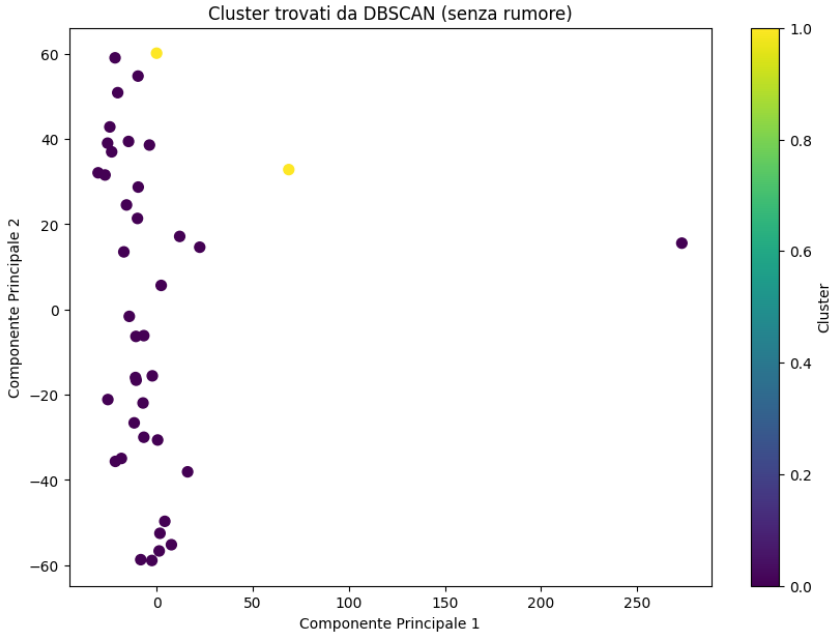

Figure 5.9: Cluster representation with DBSCAN method.

5.2 Classification Results

As described in the previous chapter, three classification algorithms were tested: Random Forest, Support Vector Machine (SVM) and Logistic Regression, comparing their performances on the various scenarios. The calculated evaluation metrics include the Area Under the Curve (AUC) and the accuracy on the test set. To complete the results, there are ROC curve plots, confusion matrices and classification reports, for each method and for each scenario, to better observe the differences of each one.

5.2.1 Classification performances - Scenario 'ALL'

Method	**AUC**	**Accuracy test set**
Random Forest	0.8925	0.7576
SVM	0.9340	0.7272
Logistic Regression	0.7431	0.5151

Table 5.4: Values of performance indicators for the three classification methods referred to the 'ALL' scenario.

In the "ALL" Scenario, SVM achieved the best AUC (0.9340), while Random Forest showed a good balance between accuracy (0.7576) and AUC (0.8925). Logistic Regression, unlike the others, provided worse results, with an AUC of 0.7431 and an accuracy of 0.5151.

5.2.2 Plot ROC Curve, Confusion Matrix and Classification Report - Scenario 'ALL'

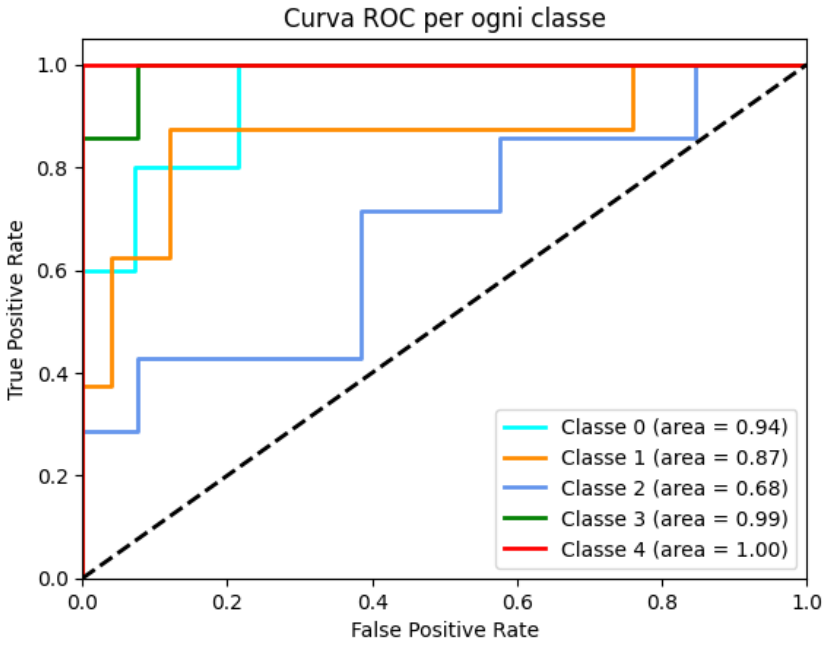

Figure 5.10: ROC Curve - Random Forest

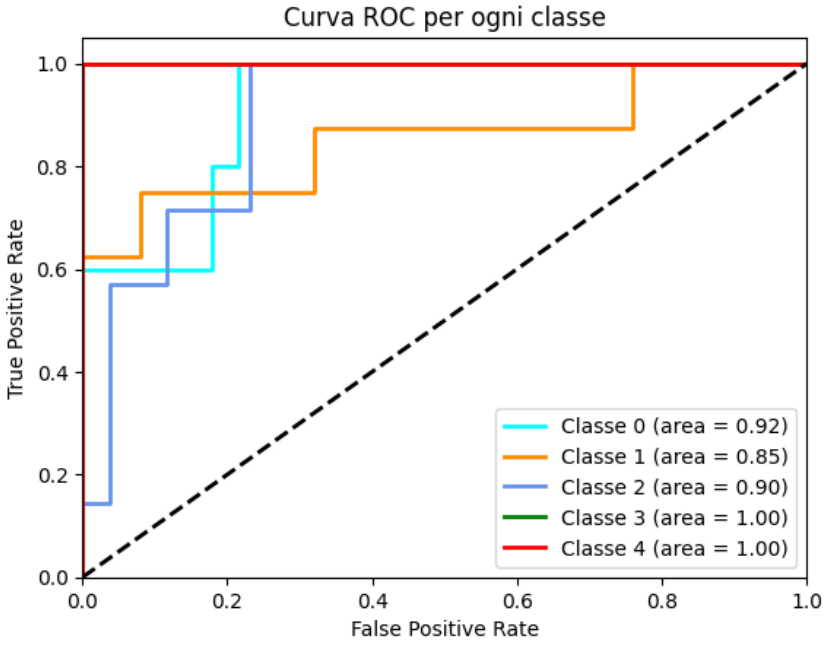

Figure 5.11: ROC Curve - Support Vector Machine

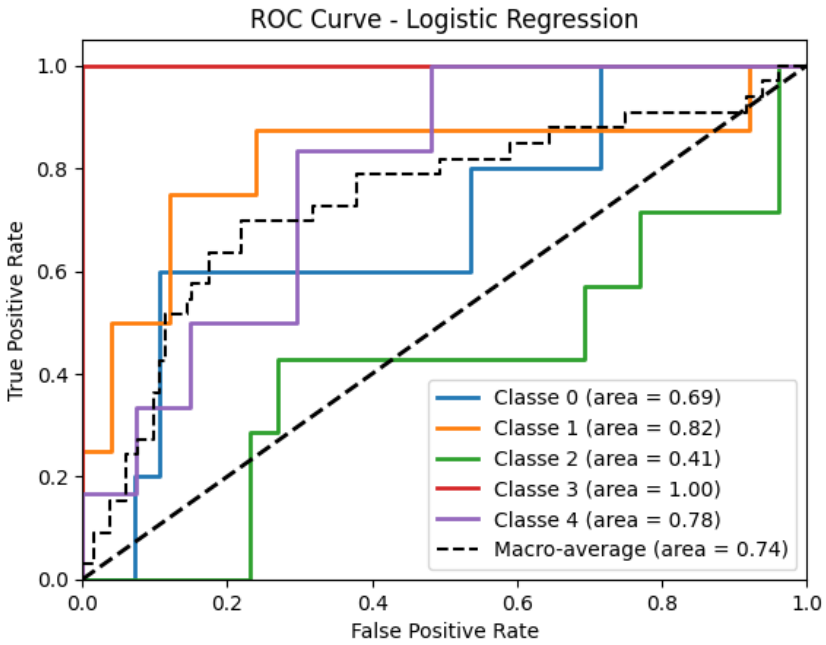

Figure 5.12: ROC Curve - Logistic Regression

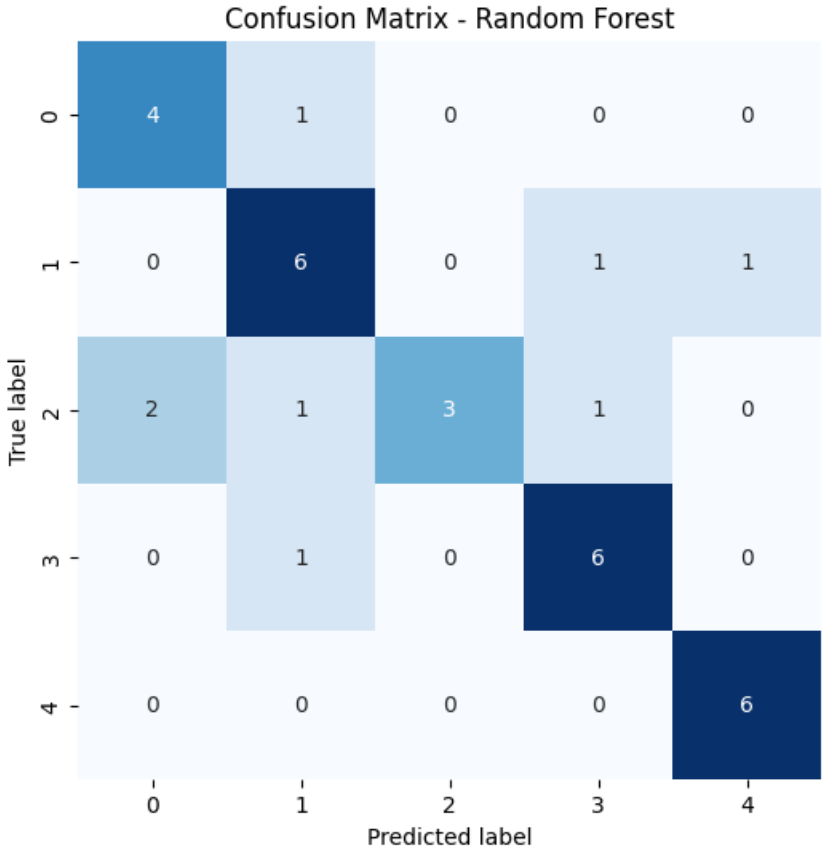

Figure 5.13: Confusion Matrix - Random Forest

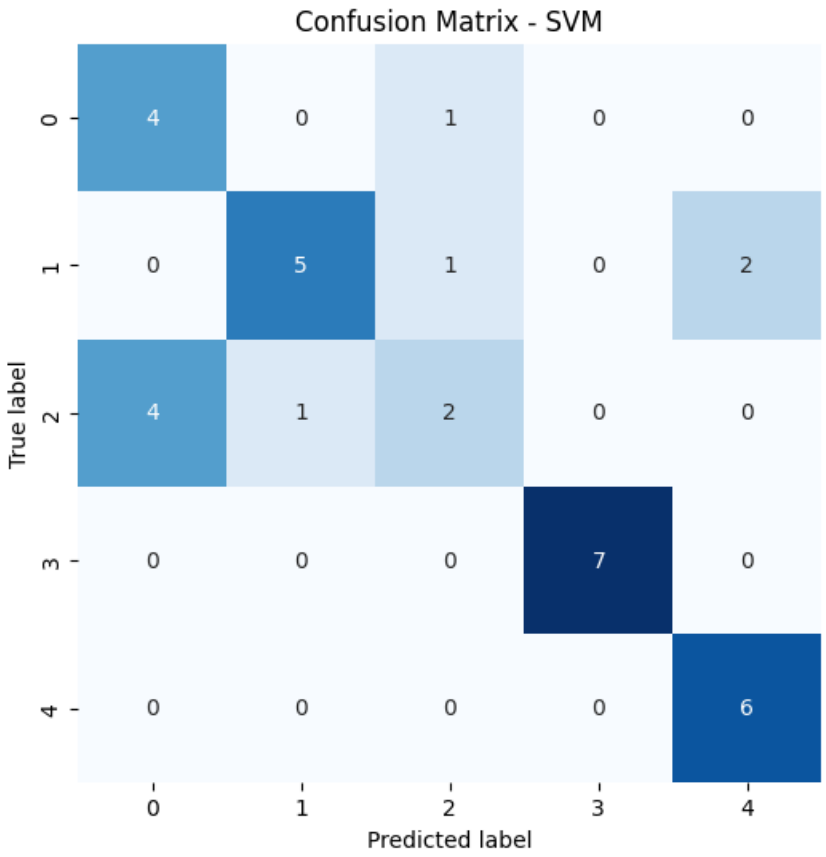

Figure 5.14: Confusion Matrix - Support Vector Machine

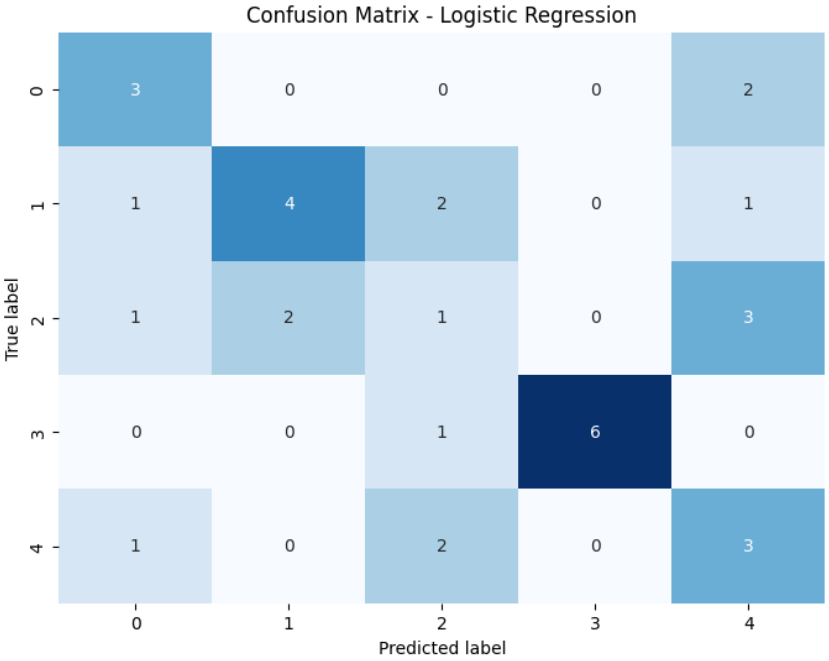

Figure 5.15: Confusion Matrix - Logistic Regression

Report di classificazione:

	precision	recall	f1-score	support
0	0.67	0.80	0.73	5
1	0.67	0.75	0.71	8
2	1.00	0.43	0.60	7
3	0.75	0.86	0.80	7
4	0.86	1.00	0.92	6
accuracy			0.76	33
macro avg	0.79	0.77	0.75	33
weighted avg	0.79	0.76	0.75	33

Figure 5.16: Classification Report - Random Forest

Report di classificazione per SVM:

	precision	recall	f1-score	support
0	0.50	0.80	0.62	5
1	0.83	0.62	0.71	8
2	0.50	0.29	0.36	7
3	1.00	1.00	1.00	7
4	0.75	1.00	0.86	6
accuracy			0.73	33
macro avg	0.72	0.74	0.71	33
weighted avg	0.73	0.73	0.71	33

Figure 5.17: Classification Report - Support Vector Machine

Report di classificazione per Logistic Regression:

	precision	recall	f1-score	support
0	0.50	0.60	0.55	5
1	0.67	0.50	0.57	8
2	0.17	0.14	0.15	7
3	1.00	0.86	0.92	7
4	0.33	0.50	0.40	6
accuracy			0.52	33
macro avg	0.53	0.52	0.52	33
weighted avg	0.55	0.52	0.52	33

Figure 5.18: Classification Report - Logistic Regression

5.2.3 Classification performances - Scenario 'OA'

Method	AUC	Accuracy test set
Random Forest	0.9811	0.8
SVM	0.9566	0.7333
Logistic Regression	0.9955	0.9333

Table 5.5: Values of performance indicators for the three classification methods referred to the 'OA' scenario.

In the "OA" Scenario, however, Logistic Regression outperformed the other two methods with a very high AUC (0.9955) and an accuracy of 93.33%. Random Forest and SVM showed AUCs of 0.9811 and 0.9566, respectively, with slightly lower accuracies.

5.2.4 Plot ROC Curve, Confusion Matrix and Classification Report - Scenario 'OA'

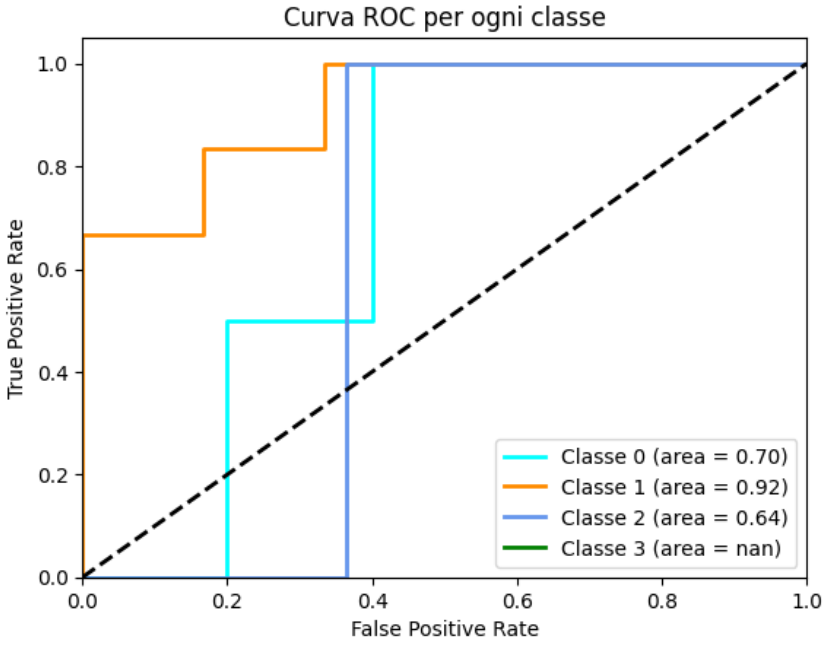

Figure 5.19: ROC Curve - Random Forest

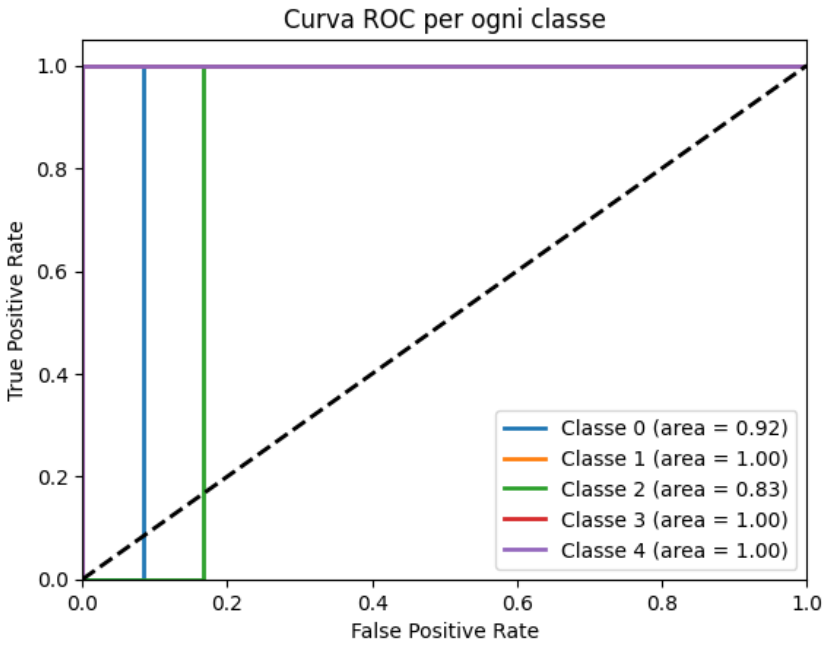

Figure 5.20: ROC Curve - Support Vector Machine

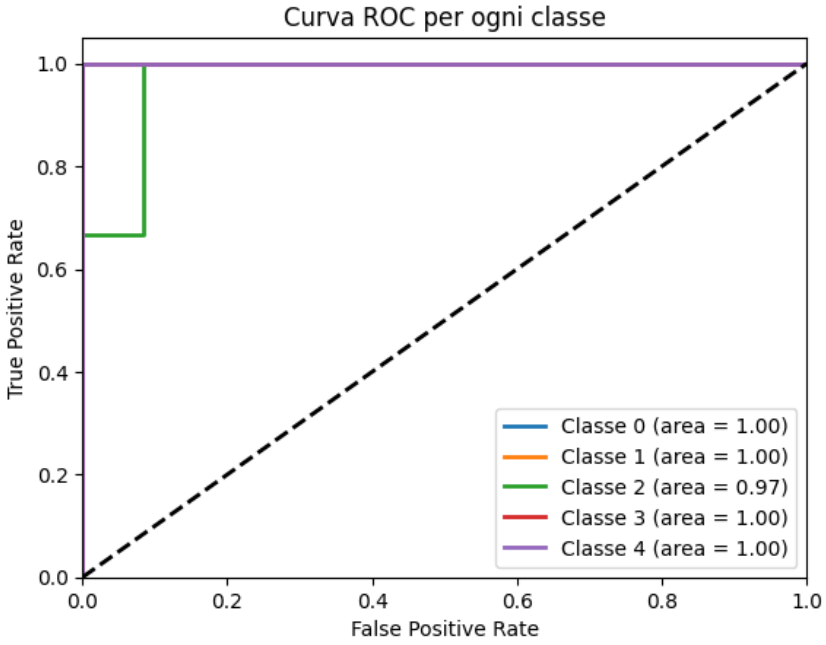

Figure 5.21: ROC Curve - Logistic Regression

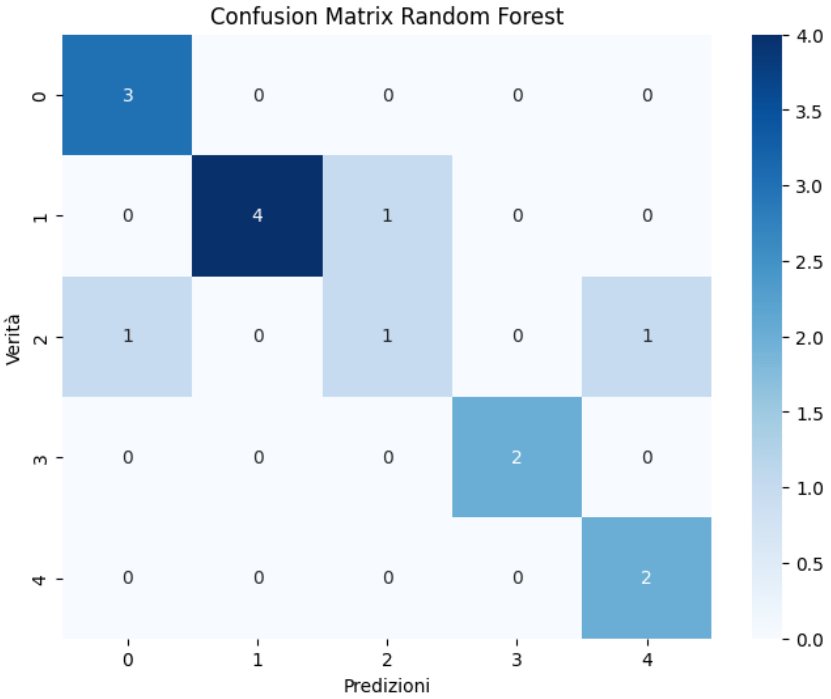

Figure 5.22: Confusion Matrix - Random Forest

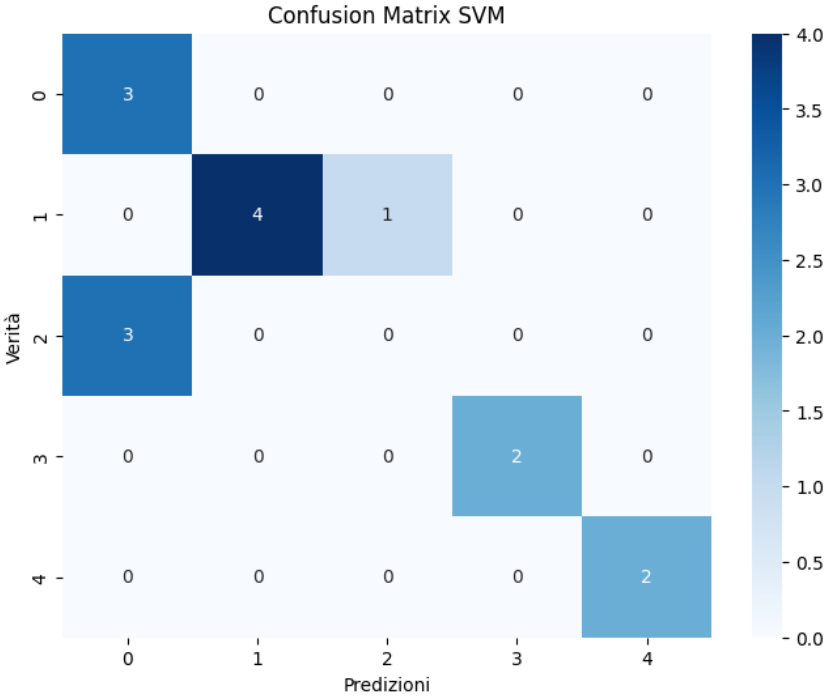

Figure 5.23: Confusion Matrix - Support Vector Machine

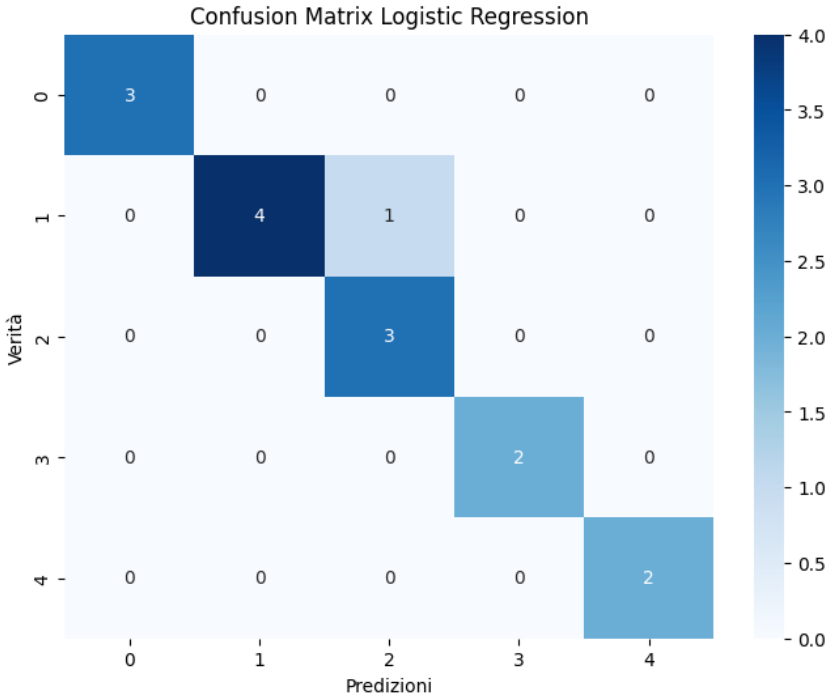

Figure 5.24: Confusion Matrix - Logistic Regression

Report di classificazione per Random Forest:

	precision	recall	f1-score	support
0	0.75	1.00	0.86	3
1	1.00	0.80	0.89	5
2	0.50	0.33	0.40	3
3	1.00	1.00	1.00	2
4	0.67	1.00	0.80	2
accuracy			0.80	15
macro avg	0.78	0.83	0.79	15
weighted avg	0.81	0.80	0.79	15

Figure 5.25: Classification Report - Random Forest

Report di classificazione per SVM:

	precision	recall	f1-score	support
0	0.50	1.00	0.67	3
1	1.00	0.80	0.89	5
2	0.00	0.00	0.00	3
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
accuracy			0.73	15
macro avg	0.70	0.76	0.71	15
weighted avg	0.70	0.73	0.70	15

Figure 5.26: Classification Report - Support Vector Machine

Report di classificazione per Logistic Regression:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	0.80	0.89	5
2	0.75	1.00	0.86	3
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	2
accuracy			0.93	15
macro avg	0.95	0.96	0.95	15
weighted avg	0.95	0.93	0.93	15

Figure 5.27: Classification Report - Logistic Regression

5.2.5 Classification performances - Scenario 'OC'

Method	**AUC**	**Accuracy test set**
Random Forest	0.925	0.6667
SVC	0.7278	0.6
Logistic Regression	0.6007	0.4666

Table 5.6: Values of performance indicators for the three classification methods referred to the 'OC' scenario.

Finally, in the "OC" Scenario, Random Forest again demonstrated good performance with an AUC of 0.925, while SVM and Logistic Regression showed poorer performance, with AUCs of 0.7278 and 0.6007, respectively. To conclude, the results indicate that the Random Forest classifier is the most balanced and effective model overall in the three scenarios in terms of AUC and accuracy.

5.2.6 Plot ROC Curve, Confusion Matrix and Classification Report - Scenario 'OC'

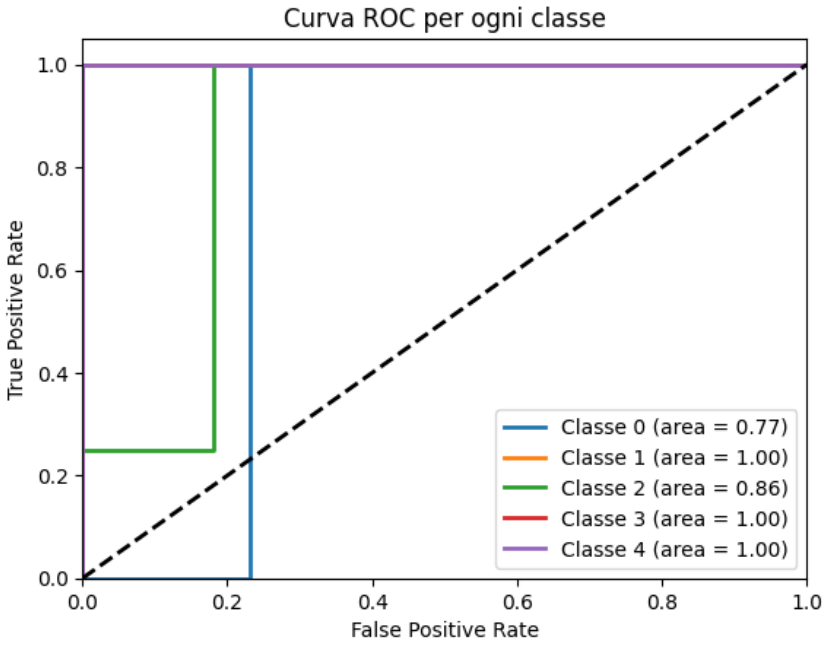

Figure 5.28: ROC curve - Random Forest

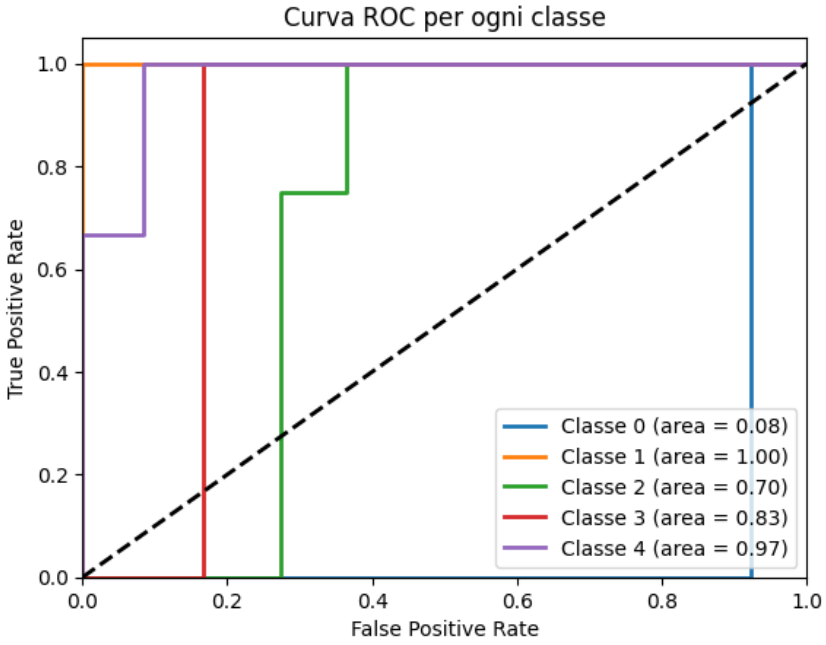

Figure 5.29: ROC Curve - Support Vector Machine

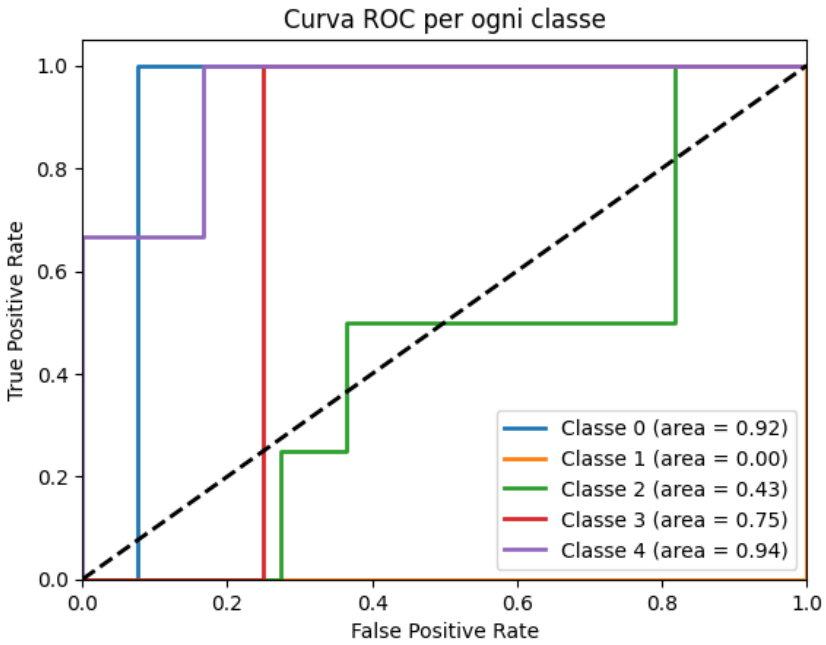

Figure 5.30: ROC Curve - Logistic Regression

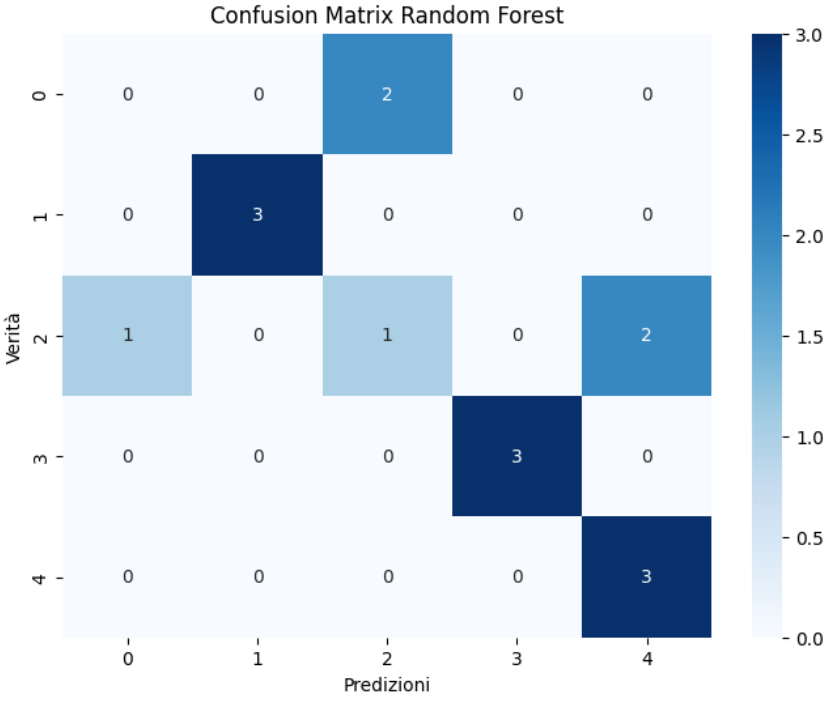

Figure 5.31: Confusion Matrix - Random Forest

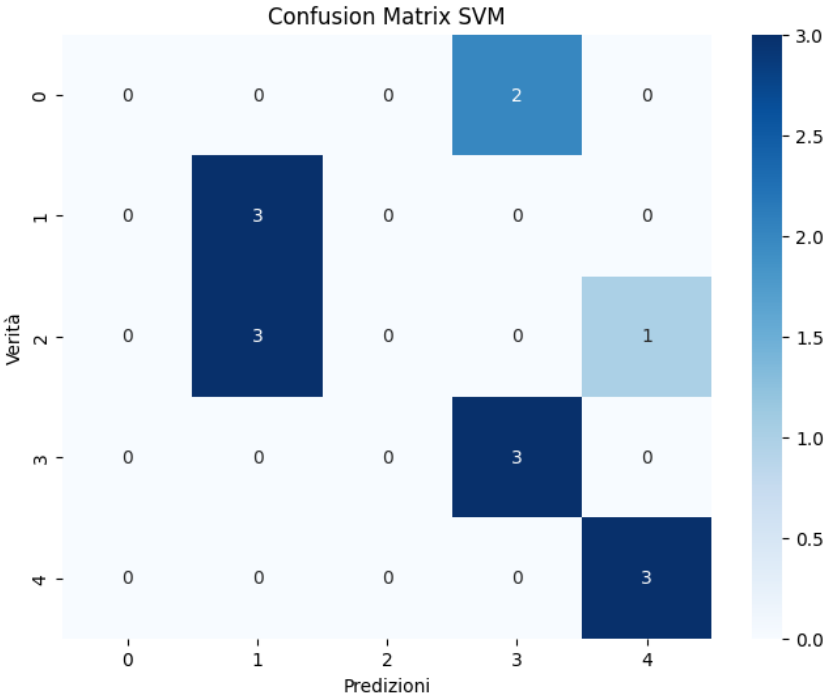

Figure 5.32: Confusion Matrix - Support Vector Machine

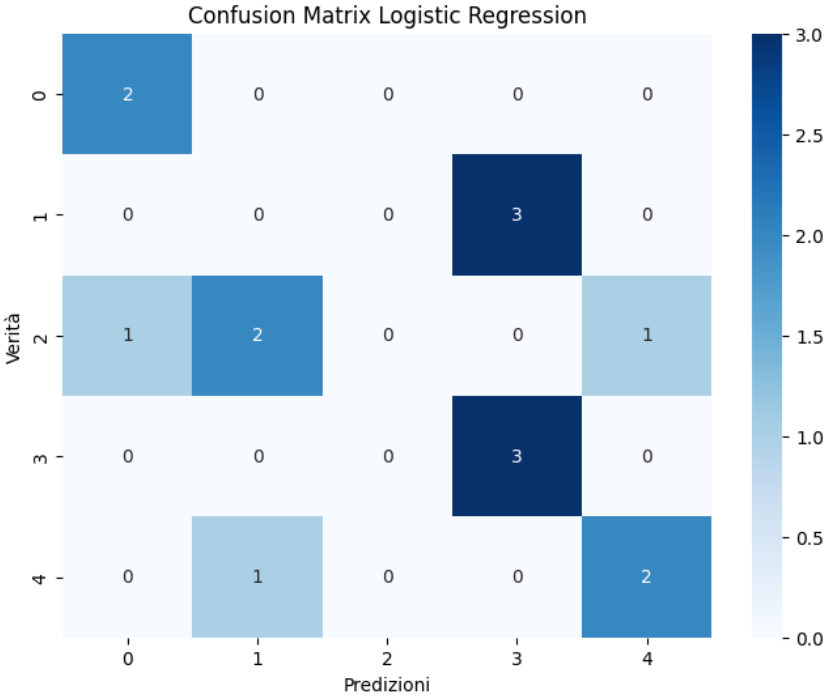

Figure 5.33: Confusion Matrix - Logistic Regression

Report di classificazione per Random Forest:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	1.00	1.00	1.00	3
2	0.33	0.25	0.29	4
3	1.00	1.00	1.00	3
4	0.60	1.00	0.75	3
accuracy			0.67	15
macro avg	0.59	0.65	0.61	15
weighted avg	0.61	0.67	0.63	15

Figure 5.34: Classification Report - Random Forest

Report di classificazione per SVM:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.50	1.00	0.67	3
2	0.00	0.00	0.00	4
3	0.60	1.00	0.75	3
4	0.75	1.00	0.86	3
accuracy			0.60	15
macro avg	0.37	0.60	0.45	15
weighted avg	0.37	0.60	0.45	15

Figure 5.35: Classification Report - Support Vector Machine

Report di classificazione per Logistic Regression:

	precision	recall	f1-score	support
0	0.67	1.00	0.80	2
1	0.00	0.00	0.00	3
2	0.00	0.00	0.00	4
3	0.50	1.00	0.67	3
4	0.67	0.67	0.67	3
accuracy			0.47	15
macro avg	0.37	0.53	0.43	15
weighted avg	0.32	0.47	0.37	15

Figure 5.36: Classification Report - Logistic Regression

Chapter 6

Discussions and Conclusions

This thesis conducted an analysis of EEG signals from patients belonging to different classes, such as healthy subjects and subjects affected by disorders of consciousness, through feature engineering and pattern recognition techniques, with the purpose of determining the different states of consciousness. The work produced interesting and promising results for research, both from the point of view of clustering and classification, showing how the use of advanced techniques can contribute to a better understanding and evaluation of EEG data and to the prediction of states of consciousness.
Examining the optimal features extracted for our three scenarios, "ALL" (all subjects), "OA" (eyes-opened subjects) and "OC" (eyes-closed subjects) showed that there are some commonalities between these scenarios, despite the contrasts in the channels on which the features were computed. In the subsets of the extracted features (see Table 6.1), it is possible to see the simultaneous presence in different groups of features based on frequency and power in different bands. In fact, in the "ALL" dataset there are features measuring the variance of power in the alpha and beta bands. Furthermore, in the "OA" dataset it is possible to discover comparable measures, despite the fact that they are correlated with the beta band, while in the "OC" the power variation in the theta band is shown. This gives us insights into the significance of the power variance in different frequency bands. Regarding Coherence and Correlation, it is possible to notice the presence of coherence variables between EEG channels both in the "ALL" and in the "OA" dataset. This leads to the argument that the relationship between brain activities recorded in pairs of channels, even if different, is an important measure both for the "OA" set and for the one in which all subjects are present.
Mutual information and Pearson correlation are, moreover, two other features that we find in more scenarios. For example, these two features are present both in the "ALL" set and in the "OC" set. This suggests that the statistical relationship between activities in different channels is significant for both measurement conditions.
Another feature, common in both the "ALL" and "OA" sets, is the delayed coherence. This measure defines the synchronization between two EEG signals with a temporal delay. The presence of Lagged Coherence in more scenarios indicates that brain interactions with temporal delay are an incisive aspect both when the eyes are open and in general acquisition conditions.
Finally, there are some features that are unique to a given single scenario. For example, the phase-locked value (PLV) appears only in the "OA" set, which may indicate that phase synchronization between EEG signals is especially important when subjects have their eyes open.

Below are the results of the evaluation parameters of the Clustering methods and also the plots of the cluster distribution.

ALL	**OA**	**OC**
mean_ch7	skew_MNF_ch8	var_theta_power_ch16
skew_MNF_ch2	var_beta_power_ch5	Skew_Std_fd_Ch14
var_alpha_power_ch5	skew_CrossCorr_9_16	skew_MI_5_15
skew_CrossCorr_13_11	skew_CrossCorr_12_4	skew_MI_15_5
skew_Coherence_15_13	skew_CrossCorr_16_9	
avg_Coherence_16_8	skew_Coherence_3_12	
skew_Coherence_16_12	skew_Coherence_9_12	
var_MI_4_17	skew_Coherence_12_9	
var_MI_10_17	skew_Pearson_9_16	
var_Pearson_9_11	avg_Pearson_16_17	
var_Pearson_11_9	skew_PLV_3_12	
skew_Pearson_11_16	skew_PLV_3_13	
skew_Pearson_16_17	skew_PLV_9_11	
skew_Pearson_17_16	skew_PLV_12_3	
skew_LC_11_13	skew_PLV_14_10	
skew_LC_13_11	skew_LC_8_16	
	skew_LC_9_16	
	skew_LC_12_3	
	skew_LC_16_8	
	skew_LC_16_9	

Table 6.1: Features extracted in different analysis scenarios.

Clustering methods, on the other hand, showed a good ability to group data based on the extracted characteristics of EEG signals, with KMeans and Agglomerative Clustering standing out for their performance in the various scenarios. However, the DBSCAN method had more difficulties in handling the changeability within the clusters, especially in contexts with variable densities in the data.
In detail, for the 'ALL' scenario, a similarity in the results in terms of Silhouette Score was observed between the KMeans and Agglomerative Clustering algorithms. Despite this, the use of the former is preferable, as it managed to obtain a separability of the data into 5 clusters equivalent to the number of classes in which the subjects were classified a priori by the medical staff. The DBSCAN, instead, not having the need to set the hyperparameter *n_clusters*, divided the data into only two groups. For the 'OA' scenario, all three methods showed very poor results, slightly better for KMeans in terms of Silhouette Score and Davies-Bouldin Index. Also in this case the number of clusters identified by DBSCAN was equal to two. Finally, for the third scenario, KMEans again provided higher results in terms of Davies-Bouldin Index and Calinski-Harabasz Index; while Allomerative Clustering provided a slightly higher value of Silhouette Score. The DBSCAN algorithm, on the other hand, struggled significantly in distinguishing clusters.
Regarding classification, Random Forest provided globally more balanced results regarding accuracy and the ability to distinguish the different classes.
In detail, for the 'ALL' scenario, higher performances were achieved with the Support Vector Machine classifier, in terms of AUC, while considering the Accuracy on the test set, Random Forest was slightly better. In the second scenario, the Logistic Regression

algorithm significantly stood out compared to the other two methods. Finally, for the third scenario, the only one to provide good performance was Random Forest.
Despite the promising results, the work presents some limitations that could be addressed in future research. The dataset showed an imbalance in the classes, with some of them represented by a too small number of samples. This could have affected the performance of the classification models, despite the use of techniques to rebalance the data.
The DBSCAN method highlighted difficulties in handling clusters with different densities, suggesting that more robust or mixed clustering methods could be explored to improve the quality of the clustering.
The feature selection, despite already efficient with the SelectKBest method, could be improved by exploring more advanced dimensionality reduction techniques, such as PCA or machine learning-based methods, for a better understanding of the most informative features of the EEG signals.
To further improve the work, several future approaches can be considered such as augmenting the dataset with the inclusion of a larger number of samples, especially for the less represented classes, and that would allow the models to generalize better and lead them to improve the overall performance.
Furthermore, the use of more complex algorithms such as deep neural networks or ensemble machine learning models could improve the predictive power, especially in the presence of noisy or more complex signals.
Also, the incorporation of temporal information such as Recurrent Neural Network (RNN) implementation techniques could provide a more dynamic view of EEG signals, improving the precision in the predictions of levels of consciousness.
In conclusion, the obtained results confirm the validity of the adopted approaches, but also showed the need for further developments to address the challenges related to the complexity of EEG signals and the variability of the classes. If sufficiently improved, the proposed techniques could significantly contribute to the development of research in the field of brain signal analysis and the automatic detection of consciousness phenomena.

Chapter 7

Ringraziamenti

Desidero innanzitutto esprimere la mia più profonda gratitudine al Professor Luca Mesin, il mio relatore, per avermi dato l'opportunità di lavorare su questo progetto. La sua competenza e il suo prezioso supporto sono stati fondamentali per la realizzazione di questo lavoro.

Un sentito ringraziamento va anche al co-relatore, Hossein Ahmadi, per la sua costante guida, la disponibilità e l'incalcolabile supporto fornito in ogni fase di questo percorso. La sua presenza è stata determinante per il completamento di questa tesi.

Rivolgo poi il mio più sincero e profondo ringraziamento alla mia famiglia, che mi ha sostenuto in ogni fase del cammino. Grazie per aver creduto in me anche quando io stesso facevo fatica a farlo. La vostra fiducia, il vostro incoraggiamento e, non da meno, i vostri sacrifici sono stati essenziali per affrontare questo percorso con più serenità.

Un grazie speciale va a mia zia Antonella, il mio "medico di fiducia" sempre presente, sempre pronta a sostenermi con i suoi preziosi consigli, soprattutto nei momenti di maggiore stress e difficoltà.

Un grande ringraziamento va allo zio Grande, per l'affetto e la stima che dimostra costantemente nei miei confronti, e per il suo supporto e i suoi consigli sempre preziosi.

Un grazie sincero anche a tutti i miei parenti: Gianfranco, Alessio, zia Annamaria, Sara, Roberta e tutti i miei cuginetti che ogni giorno dimostrano il bene che provano nei miei confronti.

Un pensiero affettuoso va anche ai miei nonni e agli zii ch purtroppo, non hanno potuto assistere a gran parte dei miei traguardi, ma che con il loro amore continuano a guardarmi da lassù.

Un ringraziamento speciale va ai miei amici di Torino: Giorgia, Marta, Selene, Matteo, Isabella, Paola e Francesca. Mi avete accompagnato in questa nuova avventura, facendomi sentire meno spaesato in una città che all'inizio mi sembrava estranea. Grazie per aver reso le giornate più leggere, per le risate e i momenti di serenità che sono stati un rifugio prezioso nei periodi più intensi. Non dimenticherò mai i pranzi della domenica, le cene del sabato sera, le abbuffate di sushi fino allo sfinimento, il fantastico weekend in Liguria

e tutte le giornate trascorse insieme.

Un ringraziamento particolare va a Paolo, con cui condivido la passione per la cucina, che tante volte ci ha fatto scherzare sull'idea di abbandonare gli studi per aprire un ristorante. Compagno di partite di pallavolo e allenamenti in palestra, mi hai sempre spinto a dare il massimo, anche quando non avevo voglia di sollevare neanche una nocciolina.

Ringrazio inoltre Lorenzo, Bruno, Luca e Jack, compagni di trasferte e partite allo stadio con cui condivido la mia più grande passione, il Lecce.

Un grazie speciale va a Leonardo e Luca, i miei coinquilini in questi ultimi due anni, che sono stati i miei primi punti di riferimento qui a Torino.

A Federico, con cui condivido tanti ricordi sin dalla nostra infanzia, uno dei pochi con cui non ci siamo mai persi di vista.

A Luigi, un amico e ormai un "doc" speciale. Non dimenticherò mai i bellissimi momenti passati insieme al mare, le partite di pallavolo in cui, modestamente, ti dimostro sempre la mia superiorità, e gli infiniti tentativi di convincerti a mangiare la migliore crêpe di Porto Cesareo. E come dimenticare il concerto dei Pinguini a Napoli, dove ho quasi rischiato la vita per il terremoto!

A Camilla, un'amica speciale, che mi ha sempre dimostrato stima e affetto. Sei stata una presenza costante, pronta a rassicurarmi nei momenti di sconforto e a calmare i miei frequenti attacchi d'ansia. Sei sempre stata felice e orgogliosa di festeggiare i miei traguardi, e voglio che tu sappia quanto ti voglio bene. Ti auguro con tutto il cuore di realizzare i tuoi sogni e raggiungere i tuoi obiettivi.

Un ringraziamento speciale va a Betty. Anche se non mi hai conosciuto in uno dei miei momenti migliori, da quella notte di San Lorenzo del 2021 (quando forse avevo bevuto un po' troppo), mi sei rimasta accanto e continui a farlo. Nonostante il tuo affetto possa a volte sembrare un po' assillante, sappi che ti voglio davvero tanto bene.

Infine, un grazie a tutti i numerosi amici sparsi per tutta Italia che ho avuto la fortuna di incontrare lungo il mio cammino. Anche se non vi ho menzionato tutti, sappiate che ciascuno di voi ha contribuito, in un modo o nell'altro, al raggiungimento di questo importante traguardo.

Grazie di cuore a tutti coloro che hanno contribuito, in qualsiasi modo, alla realizzazione di questo lavoro.

Bibliography

[1] *Agglomerative Hierarchical Clustering in Python with Scikit-Learn*. en. URL: https://stackabuse.com/bytes/agglomerative-hierarchical-clustering-with-scikit-learn/ (**urlseen** 23/09/2024).

[2] A. Almojuela, M. Hasen **and** F. A. Zeiler. "The Full Outline of UnResponsiveness (FOUR) Score and Its Use in Outcome Prediction: A Scoping Systematic Review of the Adult Literature". en. **in***Neurocritical Care*: 31.1 (**august** 2019), **pages** 162–175. ISSN: 1541-6933, 1556-0961. DOI: 10.1007/s12028-018-0630-9. URL: http://link.springer.com/10.1007/s12028-018-0630-9 (**urlseen** 13/09/2024).

[3] *Antonio Damasio — From Feelings to Socio-Cultural Homeostasis - TOWARDS LIFE-KNOWLEDGE*. en-US. **december** 2019. URL: https://bsahely.com/2019/12/11/antonio-damasio-from-feelings-to-socio-cultural-homeostasis/ (**urlseen** 17/09/2024).

[4] *Coma Recovery Scale - Revised — RehabMeasures Database*. en. **september** 2020. URL: https://www.sralab.org/rehabilitation-measures/coma-recovery-scale-revised (**urlseen** 13/09/2024).

[5] *Cos'è Random Forest? — IBM*. it. **may** 2023. URL: https://www.ibm.com/it-it/topics/random-forest (**urlseen** 17/09/2024).

[6] Kavya D. *Optimizing Performance: SelectKBest for Efficient Feature Selection in Machine Learning*. en. **february** 2023. URL: https://medium.com/@Kavya2099/optimizing-performance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48 (**urlseen** 17/09/2024).

[7] *Damasio's theory of consciousness*. en. Page Version ID: 1243847981. **september** 2024. URL: https://en.wikipedia.org/w/index.php?title=Damasio%27s_theory_of_consciousness&oldid=1243847981 (**urlseen** 17/09/2024).

[8] *Dura Mater*. URL: http://biodrawing.com/Neurology_modules/NervousSystemSite/Neuroanatomy/meninges/Dura_mater.html (**urlseen** 17/09/2024).

[9] Eric R. Kandel **andothers**. *Principle of Neural Science*. English. Fifth.

[10] Alessandro Fiori. *DBSCAN: come funziona*. it-IT. **january** 2023. URL: https://flowygo.com/blog/dbscan-come-funziona/ (**urlseen** 17/09/2024).

[11] *Glasgow Coma Scale*. it. Page Version ID: 135361436. **september** 2023. URL: https://it.wikipedia.org/w/index.php?title=Glasgow_Coma_Scale&oldid=135361436 (**urlseen** 30/11/2023).

[12] *Introduzione al clustering non supervisionato con K-Means*. it. **august** 2023. URL: https://www.diariodiunanalista.it/posts/introduzione-al-clustering-non-supervisionato-con-k-means/ (**urlseen** 23/09/2024).

[13] *La regressione logistica – Intelligenza artificiale e umana. Il giusto mix tra uomo e macchina.* it-IT. URL: https://www.humai.it/semplifichiamo/la-regressione-logistica/ (**urlseen** 17/09/2024).

[14] Steven Laureys, Adrian M Owen **and** Nicholas D Schiff. "Brain function in coma, vegetative state, and related disorders". en. **in***The Lancet Neurology*: 3.9 (**september** 2004), **pages** 537–546. ISSN: 14744422. DOI: 10.1016/S1474-4422(04)00852-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S147444220400852X (**urlseen** 13/09/2024).

[15] Luca Mesin. "Estimation of Complexity of Sampled Biomedical Continuous Time Signals Using Approximate Entropy". English. **in***Frontiers in Physiology*: 9 (**june** 2018). Publisher: Frontiers. ISSN: 1664-042X. DOI: 10.3389/fphys.2018.00710. URL: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2018.00710/full (**urlseen** 23/09/2024).

[16] Rodger Malcolm Mitchell. *What is consciousness? The hard problem. And the "sensingness" solution.* en. **december** 2023. URL: https://mythfighter.com/2023/12/17/what-is-consciousness-the-hard-problem-and-the-sensingness-solution/ (**urlseen** 17/09/2024).

[17] Kjeld Møllgård **andothers**. "A mesothelium divides the subarachnoid space into functional compartments". **in***Science*: 379.6627 (**january** 2023). Publisher: American Association for the Advancement of Science, **pages** 84–88. DOI: 10.1126/science.adc8810. URL: https://www.science.org/doi/10.1126/science.adc8810 (**urlseen** 17/09/2024).

[18] Luis Fernando Nicolas-Alonso **and** Jaime Gomez-Gil. "Brain Computer Interfaces, a Review". **in***Sensors (Basel, Switzerland)*: 12.2 (**january** 2012), **pages** 1211–1279. ISSN: 1424-8220. DOI: 10.3390/s120201211. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3304110/ (**urlseen** 16/09/2024).

[19] Marie Engelene J. Obien **andothers**. "Revealing neuronal function through microelectrode array recordings". **in***Frontiers in Neuroscience*: 8 (**january** 2015), **page** 423. ISSN: 1662-4548. DOI: 10.3389/fnins.2014.00423. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285113/ (**urlseen** 17/09/2024).

[20] G D Schott. "Penfield's homunculus: a note on cerebral cartography." **in***Journal of Neurology, Neurosurgery, and Psychiatry*: 56.4 (**april** 1993), **pages** 329–333. ISSN: 0022-3050. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1014945/ (**urlseen** 15/10/2023).

[21] J. -E. Starmark, D. Stlhammar **and** E. Holmgren. "The Reaction Level Scale (RLS 85): Manual and guidelines". en. **in***Acta Neurochirurgica*: 91.1-2 (**march** 1988), **pages** 12–20. ISSN: 0001-6268, 0942-0940. DOI: 10.1007/BF01400521. URL: http://link.springer.com/10.1007/BF01400521 (**urlseen** 13/09/2024).

[22] *Support Vector Machine (SVM) Algorithm.* en-US. Section: Machine Learning. **january** 2021. URL: https://www.geeksforgeeks.org/support-vector-machine-algorithm/ (**urlseen** 23/09/2024).

[23] Mario Tudor, Lorainne Tudor **and** Katarina Ivana Tudor. "[Hans Berger (1873-1941)–the history of electroencephalography]". hrv. **in***Acta Medica Croatica: Casopis Hravatske Akademije Medicinskih Znanosti*: 59.4 (2005), **pages** 307–313. ISSN: 1330-0164.

[24] Jonathan R. Wolpaw **and** Chadwick B. Boulay. "Brain–computer interfaces". **in** *Textbook of Neural Repair and Rehabilitation: Volume 1: Neural Repair and Plasticity*: **byeditor**Gert Kwakkel **andothers**. 2 **edition**. **volume** 1. Cambridge: Cambridge University Press, 2014, **pages** 565–576. ISBN: 978-0-511-99558-3. DOI: 10.1017/CBO9780511995583.041. URL: https://www.cambridge.org/core/books/textbook-of-neural-repair-and-rehabilitation/braincomputer-interfaces/177B253548E5AAD9F10C0FAAA1947 (**urlseen** 16/09/2024).