POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Master's Degree Thesis

EEG and Deep Learning for Predicting the Outcome of Comatose Patients After Cardiac Arrest

Supervisors

Prof. Luca MESIN

Dr. Marzia DE LUCIA

Candidate

Aurora MILANACCIO

July 2025

Summary

Most patients who experience cardiac arrest remain comatose after restoration of blood circulation due to post-anoxic brain injury. Accurate outcome prediction is essential for guiding clinical management and communicating with relatives. Among the various neuroprognostication modalities, resting-state electroencephalogram (EEG) analysis is the most widely used, as it directly reflects brain activity. Currently, EEG-based clinical evaluation relies on visual interpretation, which is prone to intra- and inter-rater variability. To overcome these limitations, computer-based methods, particularly convolutional neural networks (CNNs), have recently emerged as promising alternatives.

The aim of this work is to investigate the potential of a CNN to predict post-cardiac arrest outcome from EEG data, based on Cerebral Performance Category (CPC). Outcomes were classified as favourable (FO) for CPC scores of 1 or 2 and unfavourable (UO) otherwise.

The dataset included EEG signals from 483 patients (39% FO) from the public I-CARE database, recorded within 12–24 hours after cardiac arrest. Model optimization was performed via repeated 5-fold cross-validation, combining preprocessing strategies evaluation and Bayesian hyperparameter tuning. The best model achieved a mean validation area under the receiver operating characteristic curve (AUC) of 0.844 ± 0.050 and, on an independent test set, an AUC of 0.838 with 76% balanced accuracy. Optimization led to only marginal improvements, suggesting model robustness and limited sensitivity to preprocessing or parameter variations. Similar validation and test results indicate effective generalization to unseen data.

To gain insights into the model's decision-making process, gradient-weighted class activation mapping (Grad-CAM) was used. This method highlighted the EEG segments that most contributed to the model's predictions. The analysis revealed that the network's decisions aligned with expert knowledge, indicating that it had effectively learned clinically relevant EEG patterns.

This study confirms the strong potential of CNNs to provide stable and objective outcome predictions, supporting increasing trust in artificial intelligence and encouraging its integration into routine critical care practice.

Acknowledgements

In this section, I wish to express my sincere gratitude to those who accompanied me throughout this journey offering support, friendship, expertise and guidance.

I would first like to thank my supervisors, Professor Luca Mesin and Dr. Marzia De Lucia, for giving me the opportunity to take part in this project, offering their support and valuable guidance throughout the months of work on this thesis.

A special thanks to Dr. De Lucia who welcomed me in her research laboratory in Lausanne. The experience I had in such a stimulating and international environment was a valuable opportunity for growth, both academically and personally. I would also like to warmly thank her collaborators – Jacinthe, Andria, Giovanni and Sergi – for making me feel welcome, for their kind helpfulness and for their friendship.

A heartfelt thanks to my mom and dad, who have never stopped supporting me throughout these five years. Their help, the trust they have always placed in me and their encouragement have given me the strength to face every decision and challenge with determination.

A special thank you to my sister Rachele: you are a great point of reference for me. Knowing that I can always count on you makes every step less tiring and at the same time more secure.

I would also like to give a special thanks to the rest of my family, who have always found a way to make me feel their warm affection. Having you by my side throughout this journey, always ready to offer words of encouragement and support, has been a precious comfort to me.

A loving thanks to my boyfriend Matteo. Thank you for sharing your life with me, lightening every difficult moment and amplifying the joy of every achievement. Thank you for constantly believing in me and my abilities and for encouraging me to always give my best. The strength you give me is something I will never take for granted. You are my greatest supporter, and I am yours.

Thanks to my lifelong friends, who have been a constant in every phase of my life, and to those I've met along the university journey: it has been a privilege to share this path with you.

This achievement is also a bit of you all: without your support, it wouldn't have been the same.

Ringraziamenti

In questa sezione, vorrei esprimere la mia sincera gratitudine verso chi mi ha accompagnato in questo percorso offrendo supporto, amicizia, competenza e assistenza.

Desidero ringraziare innanzitutto i miei relatori, il Professor Luca Mesin e la Dott.ssa Marzia De Lucia, per avermi dato l'opportunità di partecipare a questo progetto, offrendomi il loro sostegno e la loro preziosa guida durante i mesi di realizzazione della tesi. Un ringraziamento particolare alla Dott.ssa De Lucia per avermi accolta nel suo laboratorio di ricerca a Losanna. L'esperienza vissuta in questo contesto stimolante ed internazionale ha rappresentato per me una grande occasione di crescita, sia dal punto di vista accademico che personale. Un sentito ringraziamento va anche ai suoi collaboratori – Jacinthe, Andria, Giovanni e Sergi – grazie per avermi fatta sentire benvenuta sin dal primo giorno, per la vostra disponibilità e per la vostra amicizia.

Un grandissimo ringraziamento a mia mamma e a mio papà, che non hanno mai smesso di supportarmi in questi cinque anni. Il loro aiuto, la fiducia che hanno sempre riposto in me e il loro incoraggiamento mi hanno dato la forza di affrontare ogni scelta e ogni sfida con determinazione. Un grazie speciale a mia sorella Rachele: sei un mio grande punto di riferimento. La certezza di poter sempre contare su di te rende ogni passo meno faticoso e allo stesso tempo più sicuro.

Grazie anche al resto della mia famiglia, che ha sempre trovato il modo di farmi sentire il proprio affetto. Sapervi al mio fianco in questo percorso, sempre pronti a offrirmi parole di incoraggiamento e sostegno, è stato per me un conforto prezioso.

Grazie di cuore al mio ragazzo Matteo. Grazie per condividere la tua vita con me, alleggerendo ogni momento difficile e amplificando la gioia di ogni traguardo raggiunto. Grazie per credere costantemente in me e nelle mie capacità, per spronarmi a dare sempre il meglio di me. La forza che mi trasmetti è qualcosa che non darò mai per scontato. Sei il mio più grande sostenitore, e io la tua.

Grazie agli amici di sempre, che rappresentano per me una certezza in ogni fase della vita, e a quelli incontrati lungo il cammino universitario: è stato un privilegio poter condividere questo viaggio con voi.

Questo traguardo è anche un po' di tutti voi: senza il vostro sostegno, non sarebbe stato lo stesso.

Table of Contents

List of Figures

List of Tables

Acronyms

A/D Analogue-to-Digital

Adam Adaptive Moment Estimation

ADC Apparent Diffusion Coefficient

ANN Artificial Neural Network

ATP Adenosine Triphosphate

AUC Area Under Curve

BCE Binary Cross-Entropy

BiLSTM Bidirectional Long Short-Therm Memory

CA Cardiac Arrest

CAR Common Average Referencing

CBF Cerebral Blood Flow

CNN Convolutional Neural Network

CPC Cerebral Performance Category

CPR Cardiopulmonary Resuscitation

CT Computed Tomography

CV Cross-Validation

DC Direct Current

DL Deep Learning

DWI Diffusion Weighted Imaging

EEG Electroencephalogram

FC Fully Connected

FN False Negative

FO Favourable Outcomes

FP False Positive

GCS Glasgow Coma Scale

Grad-CAM Gradient-Weighted Class Activation Mapping

GWR Grey-to-White Ratio

I-CARE International Cardiac Arrest Research

ICA Independent Component Analysis

ICP Intracranial Pressure

LFP Local Field Potential

MCC Matthews Correlation Coefficient

ML Machine Learning

MRI Magnetic Resonance Imaging

mRS modified Rankin Score

NPV Negative Predictive Value

NSE Neuron-Specific Enolase

PPV Positive Predictive Value

ReLU Rectified Linear Unit

ROC Receiver Operating Characteristic

ROSC Return Of Spontaneous Circulation

SSEPs Short-latency Somatosensory Evoked Potentials

TN True Negative

TP True Positive

TPE Tree-Structured Parzen Estimator

TTM Targeted Temperature Management

UO Unfavourable Outcomes

VGG Visual Geometry Group

WLST Withdrawal of Life-Sustaining Treatment

Chapter 1

Introduction

Cardiac arrest (CA) is the third leading cause of death in Europe [1]. Most patients who are successfully resuscitated from cardiac arrest, do not regain consciousness immediately after return of spontaneous circulation (ROSC) and remain comatose for a variable period, ranging from hours to weeks. Approximately half of them never regain consciousness [2]. An accurate and early prediction of neurological outcome is crucial both to communicate with the patient's relatives and to guide clinical decision-making. It helps healthcare providers to allocate resources more appropriately, focusing intensive care on patients with an higher likelihood of recovery. Among neuroprognostic tools, electroencephalography (EEG) is the most widely used, as it directly reflects brain activity. Currently, EEG is interpreted visually by expert neurophysiologists through the identification of malignant patterns. However, over the past decade, computer-based methods for predicting neurological outcome from EEG data have gained increasing attention being objective and automated. Deep learning models, particularly Convolutional Neural Networks (CNN), have shown promise in automatically extracting relevant features from raw EEG signals, potentially outperforming traditional methods. These approaches could complement clinical assessment, supporting faster and more standardized decision-making in critical care.

1.1 Cardiac Arrest

Cardiac arrest is defined as the sudden loss of all heart activity due to malfunctions in the electrical system with resulting absence of bodily blood circulation. The most frequent life-threatening cardiac arrhythmias are ventricular tachyarrhythmias (ventricular tachycardia and ventricular fibrillation), or less commonly, bradyarrhythmias, asystole and pulseless electrical activity [3].

These potentially lethal arrhythmias are usually triggered by structural heart

disease combined with functional variations or, more rarely, by functional alterations alone [4]. Structural abnormalities may include coronary artery disease, acute or healed myocardial infarction, cardiomyopathy, valvular heart disease and congenital ion channelopathy (e.g. Brugada syndrome, long QT syndrome). Functional alterations can be either cardiac, directly affecting the heart's pumping ability, or non-cardiac, which indirectly impact cardiac function. Cardiac alterations include acute myocardial ischaemia, cardiac tamponade and trauma. Non-cardiac alterations encompass electrolyte and metabolic disturbances, respiratory failure and hypoxia, autonomic nervous system dysfunctions, immunological disorders, toxicological poisoning and septic or haemorrhagic shock [5].

When CA occurs, the patient is characterized by loss of consciousness, unresponsiveness to stimuli, absence of a palpable pulse and either abnormal respiration, referred to as agonal breathing, or no respiration at all. If this condition is left untreated it can rapidly lead to death.

The actions that increase the chances of survival for a victim of CA are known as the Chain of Survival. Successful resuscitation depends on four key steps:

- Early recognition of CA and activation of the emergency medical system through a help call.
- Early bystander cardiopulmonary resuscitation (CPR) to slow down brain and heart deterioration, buying valuable time for defibrillation.
- Early defibrillation, if the rhythm is shockable (ventricular tachyarrhythmias), to restore a perfusing rhythm. If the CA occurs out of hospital, this can be achieved thanks to automatic external defibrillators placed in public spaces. Every minute of delay to defibrillation reduces the probability of survival to hospital discharge by 10-12%. However, when bystander CPR is provided, the decline in survival is more gradual, averaging a 3-5% reduction per minute of delay [6].
- Early advanced life support and standardised post-resuscitation care to restore patient's quality of life.

The chain emphasizes the interconnection of these steps and the need for each step to be performed quickly and effectively in order to optimise the chances of survival with minimal neurological impairment [7].

1.1.1 Brain injury mechanisms

The consequence of cardiac arrest is the cessation of oxygen delivery to all vital organs. Although brain represents only 2% of body weight, it requires 15-20% of total cardiac output to sustain homeostasis [8]. Brain tissue viability strongly

relies on continuous supply of oxygen and energy substrates, specifically glucose, and cessation of cerebral blood flow (CBF) results in an immediate disruption of brain activity. Due to their lack of intrinsic energy stores, neurons are particularly vulnerable to ischaemia and cellular damage starts immediately after CA [9].

Brain injury involves a complex and incompletely understood sequence of mechanisms. This process can be divided into 4 sequential, yet sometimes overlapping, phases that correspond to different stages of disease progression and treatment [6]:

1. Ischaemic depolarization;
2. Reperfusion repolarization;
3. Dysregulation;
4. Recovery and repair.

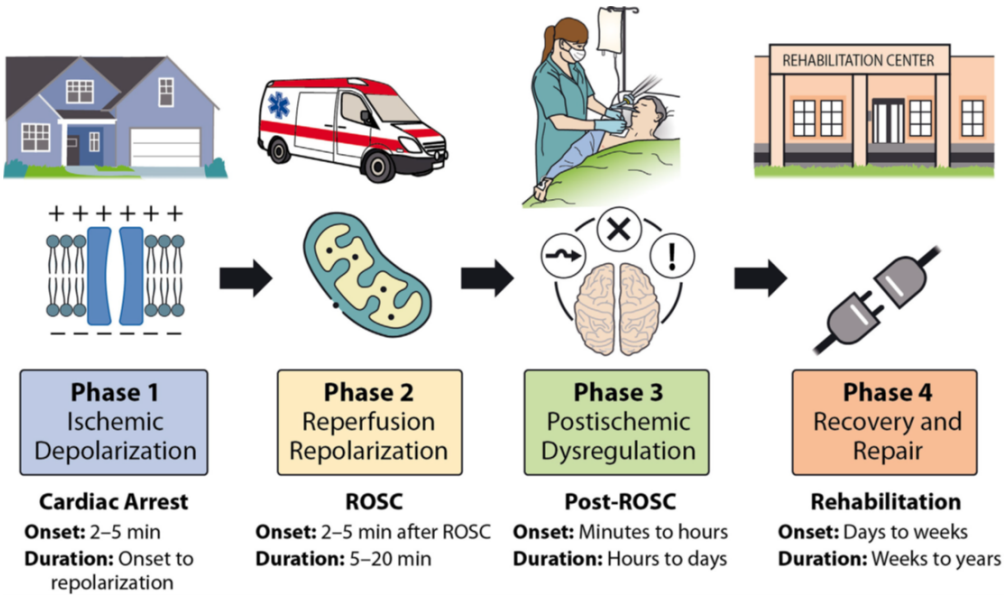

Figure 1.1: Post-cardiac arrest brain injury phases (taken from [6]).

Ischaemic depolarization

At the cellular level, ischaemia disrupts aerobic metabolism, resulting in the depletion of adenosine triphosphate (ATP), the primary high-energy substrate. The lack of ATP impairs various cellular processes. Specifically, the Na^+/K^+ pumps, which normally maintain the balance of ions across the cell membrane by actively pumping sodium (Na^+) out of the cell and potassium (K^+) into the

cell, stop working. As ATP is depleted, sodium accumulates inside the cell and potassium leaks out. The flow of sodium within the cell causes water to enter by osmosis, leading to intracellular cytotoxic oedema. Less negative intracellular environment triggers membrane depolarization, which in turn causes the opening of voltage-sensitive calcium (Ca^{2+}) channels. High concentration of intracellular calcium activates lytic enzymes that damage the cell and leads to the release of glutamate that binds to the cell membrane causing further calcium inflow. Lastly, with aerobic metabolism halted, the cell relies on anaerobic glycolysis, producing lactate, carbon dioxide (CO_2) and hydrogen ions (H^+), which leads to a decrease in pH [6], [9].

Ischaemic depolarization occurs within 2 to 5 minutes after CA [10], [11]. The CBF required to reverse ischaemic depolarization is greater than the threshold ($<$ 20% of normal CBF) at which it initially occurs and increases the longer it is left untreated [12]. Early CPR has the potential to delay ischaemic depolarization. However, since the CBF achieved by CPR ($\sim$25% of normal brain blood flow) is usually insufficient to reverse ischaemic depolarization, it cannot fully prevent neuronal damage [6].

Reperfusion injury

With ROSC, CBF is restored and, even if it is essential to ensure neuronal integrity, reperfusion of ischaemic cerebral tissue triggers mechanisms that lead to secondary brain injury. The restoration of mitochondrial electron transport chain hyperpolarizes the inner mitochondrial membrane, which enhances calcium uptake through the (Ca^{2+}) uniporter [13]. Mitochondrial calcium accumulation interferes with ATP synthesis and causes the production of reactive oxygen species, which is worsened by the oxygen supply restored by reperfusion; moreover, it can trigger the opening of mitochondrial permeability transition pore. Its permanent opening causes failure of cellular energy production potentially leading to cell necrosis, while its transient opening provokes the release of apoptosis-inducing factors [14]. Reperfusion repolarisation occurs within 2-5 minutes after ROSC and calcium overload resolves in 15-20 minutes [15], [16].

Moreover, reperfusion following prolonged global ischaemia can be incomplete and uneven due to microvascular thrombosis, endothelial oedema or neutrophil traps. In areas affected by this phenomenon, known as no-reflow, ischaemia persists instead of being reversed [6].

Dysregulation

The dysregulation phase begins within minutes to hours after ROSC and can last for hours or days. This phase is driven by multiple interconnected mechanisms including post resuscitation brain tissue hypoxia, excitotoxicity, mitochondrial

dysfunction, pathogenic inflammation and microvascular dysfunction. To these cellular processes, regional and global changes in perfusion and oxygenation are added [6].

After ROSC, a transient global hyperaemia, where CBF increases for approximately 15–30 minutes, can be observed. However, this is sometimes followed by delayed hypoperfusion, a reduction in blood flow that may contribute to secondary brain injury [8]. This phenomenon can be attributed to several factors, including altered cellular metabolism, microvascular obstruction, increased intracranial pressure due to cerebral oedema and ischaemic brain damage. At the cellular level, a key contributor to neuronal injury is excitotoxicity. The excessive release of glutamate into the extracellular space, combined with impaired reuptake by glial cells, leads to neuronal hyperexcitability. This state can manifest as epileptic seizures or abnormal electrical activity. The result is secondary intracellular and mitochondrial calcium overload, contributing to delayed neuronal death [6].

Another injury mechanism is related to the immunopathological response to damaged brain tissue. Resident macrophages, known-as microglia, are activated and secrete pro-inflammatory cytokines, attracting leukocytes from the bloodstream. These leukocytes adhere to endothelial cells of cerebral vessels and migrate into neuronal tissue thanks to increased permeability of the blood-brain barrier. This same process also allows for the leakage of fluids, leading to vasogenic oedema. The inflammatory response is further amplified by additional cytokines release from leukocytes [9]. While some degree of immune activation is necessary for tissue repair, an excessive or maladaptive immune response can worsen brain injury instead of promoting recovery [6].

Recovery and repair

Brain healing begins within days after initial brain injury and persists for weeks, months or even longer. Awakening and cognitive recovery are the most evident signs, but the underlying healing mechanisms in cardiac arrest survivors remain poorly understood and are mainly derived from knowledge of post-stroke brain remodelling [17]. The two key compensatory processes through which the brain reshape its neural networks and functional connectivity are neuroplasticity and neurogenesis. Neuroplasticity is the brain's ability to reorganize itself by forming new connections between regions. This occurs through synaptic pruning, where dysfunctional synapses are lost, and synaptic sprouting, during which new axons and dendrites grow from existing neurons [18]. Neurogenesis, on the other hand, refers to the generation of new neurons from endogenous stem cells and it is usually matched with angiogenesis and neuroglial genesis. These mechanisms can be positively influenced by pharmacological or non-pharmacological interventions, such as rehabilitation, exercise and sleep [19].

1.1.2 Neuroprotective interventions

Neuroprotective management after cardiac arrest aims to maintain physiologic homeostasis and minimise secondary brain injury. Disturbances in oxygenation, ventilation, blood pressure and temperature should be prevented [20].

Normal oxygenation should be maintained in order to prevent both hypoxaemia, which may worsen brain ischaemia, and hyperoxaemia, which could increase the production of free radicals [21].

Low partial pressure of carbon dioxide ($PaCO_2$), or hypocapnia, is associated with vasoconstriction, potentially reducing cerebral blood flow and worsening cerebral ischaemia. On the other hand, hypercapnia causes vasodilation and may increase intracranial pressure (ICP), especially in the presence of cerebral oedema. However, mild hypercapnia, which leads to a moderate increase in CBF, could be beneficial [9]. Since there is no strong evidence to suggest that mild hypercapnia is superior to normocapnia, current guidelines recommend maintaining normal $PaCO_2$ levels and avoiding hypocapnia. Additionally, hypoperfusion should be avoided in order not to reduce CBF and increase the risk of ischaemic damage [20].

Another neuroprotective strategy is targeted temperature management (TTM), which aims to maintain a specific body temperature, typically between $33°C$ and $37°C$. While experimental models have demonstrated that hypothermia ($32°C - 34°C$) can slow down cerebral metabolism and reduce brain damage, clinical studies have yielded conflicting results [20]. Due to the lack of conclusive evidence on the effectiveness of active cooling, latest guidelines no longer recommend a specific target temperature but emphasize the importance of avoiding fever ($> 37.7°C$) for at least 72 h, as hyperthermia is associated with worse neurological outcomes [22]. Despite these recent findings, therapeutic hypothermia has been and still remains part of clinical practice in some centres. In these cases, a fixed temperature is typically maintained for 24 h, followed by a gradual rewarming phase.

In addition to the aforementioned strategies, pharmacological interventions may also play a role in neuroprotection by mitigating alterations in the pathways triggered by cardiac arrest; however, their efficacy requires further clinical evaluation. [20].

1.2 Neuroprognostication

Approximately two-thirds of deaths in comatose patients admitted to the intensive care unit, after resuscitation from out-of-hospital-cardiac-arrest, are caused by hypoxic-ischaemic brain injury [23], [24]. However, only in a minority of cases, death occurs as a direct consequence of post-cardiac arrest brain damage leading to irreversible loss of all brain functions, i.e. brain death [25]. Most of these deaths result from active withdrawal of life-sustaining treatment (WLST) in patients

where the severity of brain injury is such that it indicates very low probability of neurologically meaningful survival [26].While decisions regarding WLST should also consider factors such as the patient's age, comorbidities and personal wishes, neurological prognostication plays a crucial role in guiding post-arrest care [20]. In particular, early and accurate neuroprognostication is important both to inform the patient's relatives and to avoid unnecessary prolonged treatment in patients with no chance of achieving a favourable neurological outcome.

1.2.1 Outcome measures

Neurological outcome following resuscitation from cardiac arrest is most commonly assessed through the Cerebral Performance Category (CPC). This scale includes five scores, ranging from complete recovery to death (Table 1.1).

Table 1.1: Cerebral Performance Category Score (CPC) (taken from [27]).

Score	Description
1	Conscious: alert, able to work and lead a normal life. May have minor psychological or neurological deficits (mild dysphasia, nonincapacitating hemiparesis, or minor cranial nerve abnormalities)
2	Conscious: sufficient cerebral function for independent activities of daily life; able to work in a sheltered environment
3	Conscious: dependent on others for daily support because of impaired brain function (in an institution or at home with exceptional family effort). At least limited cognition. Includes a wide range of cerebral abnormalities from ambulatory with severe memory disturbance or dementia precluding independent existence to paralytic and able to communicate only with eyes, as in the locked-in syndrome
4	Not conscious: unaware of surroundings, no cognition. No verbal or psychological interactions with environment
5	Certified brain dead or dead by traditional criteria

Specifically, CPC 1 represents no or minimal neurological disability; CPC 2 minor neurological disability; CPC 3 severe neurological disability; CPC 4 coma or vegetative state and CPC 5 death. CPC 1 and 2 are universally considered as good neurological outcome, as they correspond to patients independent in daily activities, while CPC 4 and 5 are invariably associated with poor neurological outcome. Regarding CPC 3, it is generally, but not universally, considered as unfavourable neurological outcome, since it includes patients that need assistance for daily living [9]. CPC is typically assessed 3 to 6 months after CA, usually through patient

interviews or medical record reviews, however there is no standardized method for its collection.

Another scale that can be used to assess neurological functions is the modified Rankin Score (mRS) [28]. Originally developed for stroke, it was then adapted to cardiac arrest patient assessment. The mRS includes seven scores, from 0 to 6, and provides a more fine-grained assessment of favourable outcomes compared to CPC, making it a potentially more suitable tool for measuring long-term neurological recovery [9].

Neither the CPC nor the mRS differentiate between the two main causes of neurological death: brain death and death due to WLST. More importantly, they do not distinguish between neurological and non-neurological causes of death. As a result, resuscitated patients who pass away due to extracerebral complications after regaining consciousness are classified as CPC 5 or mRS 6, despite their neurological status at the time of death. To address this issue, the best neurological score, rather than the final one, during the observation period can be used [9]. It should be noted however that relying on the best score may overlook later neurological deterioration that could impact the final prognosis.

1.2.2 Prognostic predictors

Since no single predictor can provide absolute accuracy, assessing the severity of hypoxic-ischemic brain injury requires a multimodal approach (Figure 1.2).

European post-resuscitation care guidelines (2021) recommend integrating information from clinical examination, biomarkers, neurophysiology and neuroimaging into a prognostic algorithm (see subsection 1.2.4) to evaluate the probability of a poor neurological outcome in comatose patients [20].

Clinical examination

The most used clinical prognostic examination signs are motor response, ocular reflexes and myoclonus [9].

Motor response is evaluated through the motor component of the Glasgow Coma Scale (GCS), a standardized tool for measuring a patient's consciousness level. The lack of motor response or abnormal extensor or flexor response following a painful stimulus, in comatose patient after 72 h from ROSC (GCS-M ≤ 3), is a very sensitive sign of poor neurological outcome.

Ocular reflexes originate in the brainstem, which is quite resistant to anoxic injury. This makes their absence a more specific sign of severe brain injury rather than an altered motor response, which can be generated on different levels between the cortex and the brainstem. At ≥ 72 h after ROSC, the bilateral absence of both pupillary and corneal reflexes predicts poor neurological outcome.

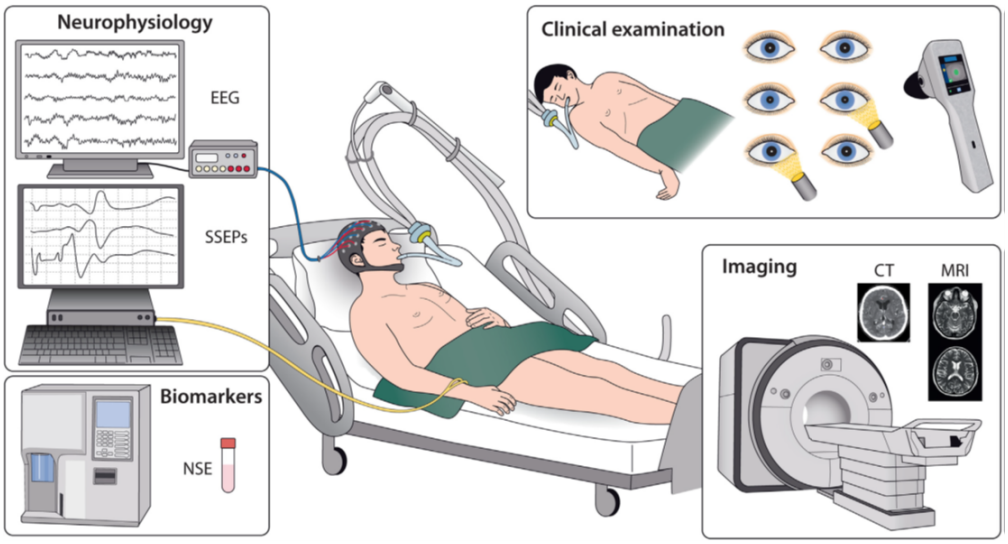

Figure 1.2: Neurological prognostication modes (taken from [20]).

Myoclonus refers to sudden, brief and involuntary twitching, jerking or spasm caused by contraction or relaxation of one or more muscles. Its presence may indicate an unfavourable prognosis depending on its characteristics. Early occurrence ($<$ 6 h after ROSC), generalised distribution, synchronous or stereotyped patterns and prolonged ($>$ 30 min) duration (status myoclonus) are associated with worse outcomes [20].

Biomarkers

After cardiac arrest, neurons and glial cells release several molecules that can be measured in serum or plasma as markers. Blood biomarkers are easy to detect and can objectively quantify the extent of brain injury. However, their interpretation can be challenging, since their value depends on laboratory instruments and protocols and/or potential extracerebral sources [9].

Among biomarkers, neuron-specific enolase (NSE) is the only one recommended for neurological prognosis [20]. NSE is a glycolytic enzyme released into the bloodstream by damaged neurons and neuroendocrine cells. This marker is also present in blood cells and its early increase in concentration can be due to haemolysis, for example, caused by CPR [29]. However, high NSE values at 48-72 h after ROSC are more likely to reflect a brain injury [20].

Neurophysiology

Short-latency somatosensory evoked potentials (SSEPs) and EEG are the two main electrophysiological methods used to assess the chances of neurological recovery after CA.

SSEPs are non-invasively recorded by stimulating the median nerve at the wrist using transcutaneous electrodes. The electrical signals, generated in the primary somatosensory cortex, are detected by electrodes placed over the contralateral skull [30]. The negative wave observed after 20 ms on the scalp EEG, which reflects activation of the somatosensory cortex, is referred to as the N20 wave. The bilateral absence of the cortical N20 wave after CA mostly indicates severe brain injury with high specificity and moderate to low sensitivity. Its detection is resistant to hypothermia but can be influenced by muscular artefacts, hence the use of myorelaxants is recommended when recording SSEPs [20]. A low amplitude of N20 wave has been shown to be a sign of poor outcome as well [31], [32].

With regards to EEG, it is the most widely used test to evaluate the extent of post-cardiac arrest brain injury in clinical practice [33]. Since this signal is the main focus of this thesis, section 1.3 will explore its role in detail, the specific patterns associated with neurological outcomes and the challenges in its interpretation.

Imaging

Brain computed tomography (CT) and magnetic resonance imaging (MRI) are valuable tools to evaluate brain injury by detecting cerebral oedema following cardiac arrest. However, acquiring CT and MRI images in the intensive care unit can be challenging due to patient instability, the need for specialized equipment and logistical constraints, which may limit their routine use in this setting.

On brain CT images, vasogenic oedema appears as effacement of cortical sulci. In contrast, neuronal swelling due to cytotoxic oedema leads to a decrease in grey matter (neurons) density, while the white matter (axons) remains relatively unaffected. As a result, the grey-to-white ratio (GWR) decreases, making the grey/white matter interface less visible. The lower the GWR, the more severe the brain oedema. A reduced GWR occurs early in patients with severe hypoxic-ischaemic brain injury and has been proposed as an objective predictor. However, its reliability is limited by high variability across studies due to differences in sampling areas and scanner software and hardware [9], [20].

Hypoxic-ischaemic brain injury reduces water diffusivity which appears on MRI as hyperintensity on diffusion weighted imaging (DWI) with corresponding low apparent diffusion coefficient (ADC) values. Similarly to GWR, ADC values vary across studies.

Since there is currently no standardized method for CT-GWR or MR-ADC measurements, these techniques are better suited to confirm the presence of ischaemic

injury through visual analysis by an experienced neuroradiologist [20].

1.2.3 Sources of bias in prognostication

One of the main sources of bias in neuroprognostication following CA is the so-called self-fulfilling prophecy. This occurs when results of prognostic tests are used to guide therapeutic decisions that contribute to the fulfillment of the predicted outcome. This leads to an overestimation of test performances and, in unfortunate cases, to inappropriate WLST. Ideally, self-fulfilling prophecy could be avoided by blinding test results. However, this is not feasible for clinical examination and, specifically for EEG or brain images results, would be unethical since they may reveal potentially treatable complications (e.g. epileptic seizures, intracranial hypertension). An obvious way to limit the confirmation bias caused by the self-fulfilling prophecy is to investigate prognostication where there is no active WLST policy.

Other strategies to reduce the risk of falsely pessimistic prognosis include not performing prognostication tests in the presence of confounding factors and basing decisions on multimodal approaches and on repeated assessments. Among prognostication methods, clinical examination is influenced by sedatives, myorelaxants and opioids, EEG patterns by hypothermia and sedatives and the N20 wave amplitude upon SSEP assessment by profound sedation.

In addition, it is necessary to consider that neurological recovery takes time and therefore an appropriate time lag between the prognostication test and the assessment of neurological outcome is needed [20]. Guidelines recommend for this to be performed between 3 and 6 months after cardiac arrest [34].

The last source of bias worth mentioning is non-neurological causes of death. In fact, when evaluating the accuracy of neurological predictors, extracerebral causes of death, that can occur after regaining consciousness, should be taken into account [9].

1.2.4 Prognostication algorithm

In the 2021 European guidelines on post-resuscitation care, a prognostication strategy to predict neurological outcome of comatose adult patients after CA has been proposed. Prognostication assessment should follow the scheme reported in Figure 1.3. Two or more concordant unfavourable signs are necessary to prognosticate poor neurological outcome. In case of discordant tests, where some indicate unfavourable outcome and others favourable outcome, a prognostic reassessment is recommended [20].

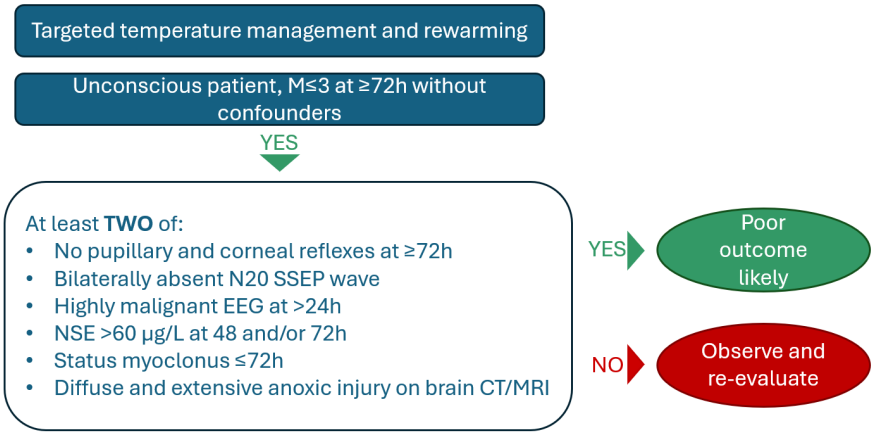

Figure 1.3: Prognostication algorithm after cardiac arrest based on current European guidelines on post-resuscitation care (adapted from [20]).

1.3 EEG in neuroprognostication

1.3.1 From neurons to EEG

EEG represents the most common way to non-invasively measure electrical brain activity. In particular, EEG measures the potential differences between two points on the scalp through surface electrodes placed in a cap.

The human brain contains 100 billion neurons, each of which is connected to thousands of other neurons. This large electrical network can be divided into many sub-networks. The activity of a sub-network causes changes in extracellular potentials; their superimposition is defined as a local field potential (LFP) [35]. The main sources of LFP are synaptic potentials, the transmission signals between neurons, and action potentials, the electrical activations of neurons in response to stimuli [36].

The potentials recorded on the scalp are a modified version of LFPs originating from underlying groups of neurons. Two main mechanisms are responsible for this phenomenon. First, the electrical field decays with the square of the distance from the source and, therefore, LFPs are attenuated when they reach the electrodes. Second, the volumetric conductance of the head's tissues causes the potential generated at a single point to spread over a wider area resulting in spatial smoothing of the original signal [35].

Due to attenuation and smoothing, only synchronous neuronal activity (summated activity across brain areas) can be measured at the scalp level [37]. Major EEG contributors are postsynaptic potentials of pyramidal neurons in the cerebral cortex, as action potentials are too brief to be recorded [38]. Since the neuronal

soma is located in deeper cortical layers, while neuronal dendrites extend more superficially, synaptic activity creates a charge distribution that makes them dipoles oriented perpendicularly to the cortical surface. If many neurons receive synaptic transmissions synchronously, the resulting electrical field magnitude is strong enough to propagate up to the surface. The larger the population of neurons involved in synchronous activity, the greater the amplitude of the signal recorded on the scalp surface. This summation is less effective in cortical sulci, where pyramidal neurons are not perpendicularly aligned to the scalp, leading to signal cancellation [37].

To sum up, the ability of EEG to record brain activity strongly depends on neuronal synchronization, the number and orientation of neurons, their distance from the surface and the electrical properties of the surrounding tissues.

The oscillations of neural activity can have predominant frequencies that reflect the underlying cognitive state. These frequency bands include delta (0.5-4 Hz) associated with deep sleep, theta (4-8 Hz) linked to drowsiness or light sleep, alpha (8-13 Hz) observed during relaxed wakefulness, beta (13-30 Hz) indicative of active thinking or focused attention and gamma (30-100 Hz) related to high-level cognitive functions such as attention and memory.

In the context of post-cardiac arrest coma, the EEG is typically dominated by slow-frequency rhythms. Gamma activity is generally absent, while beta rhythm is rare and most often pharmacologically induced. The cerebral activity observed in hypoxic-ischemic encephalopathy typically falls within the delta, theta and alpha bands, although their spatial distribution and reactivity can vary considerably across patients and over time [39].

1.3.2 EEG recording system

Brain electrical activity is primarily detected by surface electrodes, although achieving a clean and reliable signal requires addressing various challenges.

The EEG signal is inherently weak, typically ranging between 10-100 μV in a healthy, awake subjects and therefore it may be masked by other interferences, generated both by the body and the environment [40]. Physiological artefacts can include muscles activity, eye movements, heart activity and subject movements. Environmental noise mainly comes from power line interference, electrode cable movements and temporary electrode detachments (electrode pop artifact), which can increase impedance and degrade signal quality [41], [42].

To minimize these issues, EEG systems use high-gain differential amplifiers. These devices amplify the differential electrical activity between two electrodes while rejecting common signals, effectively reducing shared noise and enhancing the relevant EEG signal [41]. Additionally, careful electrode placement and proper impedance control are crucial for obtaining reliable recordings.

However, despite precautions, some noise will inevitably overlap with the neural signal. Therefore, signal processing, both analogue and digital, is essential to further enhance signal quality and extract meaningful brain activity.

Once the signal has been amplified, it is sampled at fixed time intervals, and each sample is then converted into a digital form by an analogue-to-digital (A/D) converter. The A/D converter is connected to a recording device, such as a computer, where the signal can be displayed and stored.

1.3.3 Electrode placement and montages

The most commonly used electrodes material is pure silver coated with a layer of silver chloride (Ag/AgCl) as they offer excellent and stable electrical properties.

To facilitate interpretation and comparison, electrode placement follows standardized positioning systems. One of the most widely recognized is the 10–20 system, originally proposed by Jasper [43], which ensures uniform interelectrode spacing. This system positions electrodes at intervals of 10% and 20% relative to four anatomical landmarks: the nasion, inion and left and right preauricular points (Figure 1.4). To improve spatial resolution, additional electrodes can be incorporated into the 10–20 framework. According to the convention, electrodes are named based on their position on the scalp. The first character refers to the cortical area (F=frontal area, C=central area, P=parietal area, T=temporal area and O=occipital area). The second character is a number or a letter. Odd numbers correspond to sites on the left hemisphere, while even numbers represent sites on the right hemisphere. Midline electrodes are indicated with the letter z, Additionally, the numbers increase as the distance from the midline increases [37].

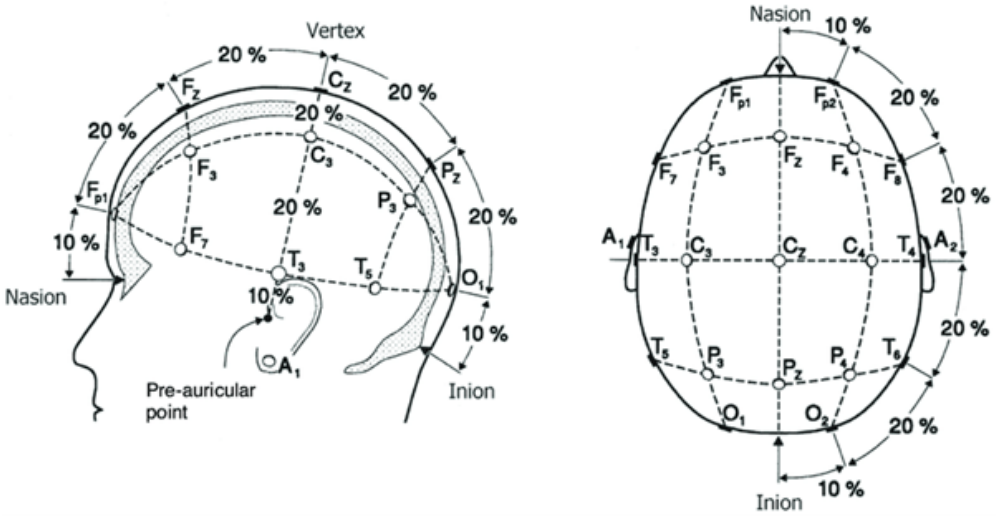

Figure 1.4: Electrode placement according to the international 10-20 system (taken from [44]).

Based on the specific points on the scalp where the potential difference is measured, there are two main types of electrode montages: unipolar (or referential) montage and bipolar montage. In the unipolar setup, the potential difference is measured relative to the same reference electrode for all recording sites. The reference is typically placed on locations such as the earlobe, nose, mastoid, chin, neck or the centre of the scalp. However, there is no universal consensus on the optimal position for the reference electrode, as bioelectric currents generated by muscles, heart or brain activity propagate throughout the body. On the other hand, in the bipolar montage, each channel registers the potential difference between two particular scalp electrodes [40].

1.3.4 Prognostic value of EEG

EEG is the most used and accessible prognostic tool following cardiac arrest [33]. This signal, reflecting cortical synaptic activity, which is very sensitive to the effects of hypoxia, provides valuable information about the gradual recovery from brain injury on time scales of hours to days. Moreover, EEG plays a crucial role in diagnosing and managing epileptiform activity [45], [46].

The evolution of EEG background and patterns in comatose patients after CA can help predict neurological outcome. In 2021, the American Clinical Neurophysiology Society proposed standardized terminology for EEG in critical care patients [47].

Three main aspects should be considered when assessing EEG: the background activity, the presence of superimposed discharges and the reactivity to stimulation [20].

Background activity

The EEG background is described according to its frequency, voltage and continuity. Signal voltage is classified as normal ($\geq 20\ \mu V$), low ($< 20\ \mu V$) or suppressed ($< 10\ \mu V$). Continuity is categorised as continuous, discontinuous, burst-suppression or suppression [45]. EEG is considered continuous if no suppression is observed throughout the recording; discontinuous if suppression periods represent 10-49% of the recording; burst-suppression if 50-99% consists of suppressed periods alternated with bursts; and suppression if the entire signal is suppressed [47].

Most patients show suppressed or low voltage EEG shortly after return to spontaneous circulation. However, in patients who later recover, a gradual transition towards normal voltage EEG is usually observed within 12-24 h from cardiac arrest. The faster this normalization occurs, the better the outcome. Persistent presence of suppressed background after 24 h from ROSC is a reliable sign of poor neurological prognosis [45]. Burst-suppression activity is widely recognized as a highly unfavourable sign, particularly if it occurs 12-24 h after ROSC [20]. Bursts

are defined as waveforms lasting more than 0.5 s and having at least 4 phases. These bursts can be further divided in highly epileptiform and identical. Highly epileptiform bursts exhibit epileptiform discharges (spikes or sharp waves) in at least half of their duration. On the other hand, bursts are defined as identical if the first ≥ 0.5 s of each burst appear visually similar in all channels and in 90% of bursts [47]. Identical bursts are a more specific sign of poor outcome than highly epileptiform bursts, as they suggest a deterministic process indicative of severe neurological impairment. A burst-suppression pattern may also be transiently induced by sedation, but bursts appear heterogeneous and transient [48].

In contrast, the prognostic value of discontinuous EEG is uncertain across studies [20].

Superimposed activities

The discharges superimposed on the EEG signal can be distinguished into periodic, sporadic epileptiform and electrographic seizures [20]. Differently from bursts, discharges are waveforms lasting less than 0.5 s, regardless the numbers of phases, or lasting more than 0.5 s and with a maximum of 3 phases.

Periodic discharges occur repeatedly with a quantifiable interdischarge interval and a relative uniform morphology [47]. They can involve both hemispheres (generalised periodic discharges) or just one of them (lateralised periodic discharges). This superimposed pattern is usually related to worse prognosis. However, the background on which periodic discharges appear is considered a stronger predictor of neurological outcome [49].

Sporadic epileptiform discharges refer to non-periodic epileptiform activity. They can be linked to unfavourable outcome, but their presence has uncertain prognostic value [50].

Electrographic seizures are defined as epileptiform discharges occurring at an average frequency of more than 2.5 Hz for at least 10 s or any EEG pattern that shows clear evolution (in terms of frequency, location or morphology) that lasts for more than 10 s. An electrographic seizure lasting more than 10 continuous minutes or 20% of one hour recording long is termed electrographic status epilepticus [47]. These patterns are generally associated with poor prognosis, especially when combined with other unfavourable EEG features. However, recovery remains possible when the EEG is continuous and reactive and the epileptiform activity is transient or responds to treatment [51].

EEG reactivity

EEG reactivity consists of a measurable change in frequency or amplitude, following a predefined stimulus (pain, auditory or light) [47]. High variability across reactivity testing has been reported, contributing to the inconsistent prognostic value of

EEG reactivity in the literature. Presence of EEG reactivity in conjunction with continuous or discontinuous normal voltage is a sign of good outcome, while its absence has little added value to the EEG background activity [52].

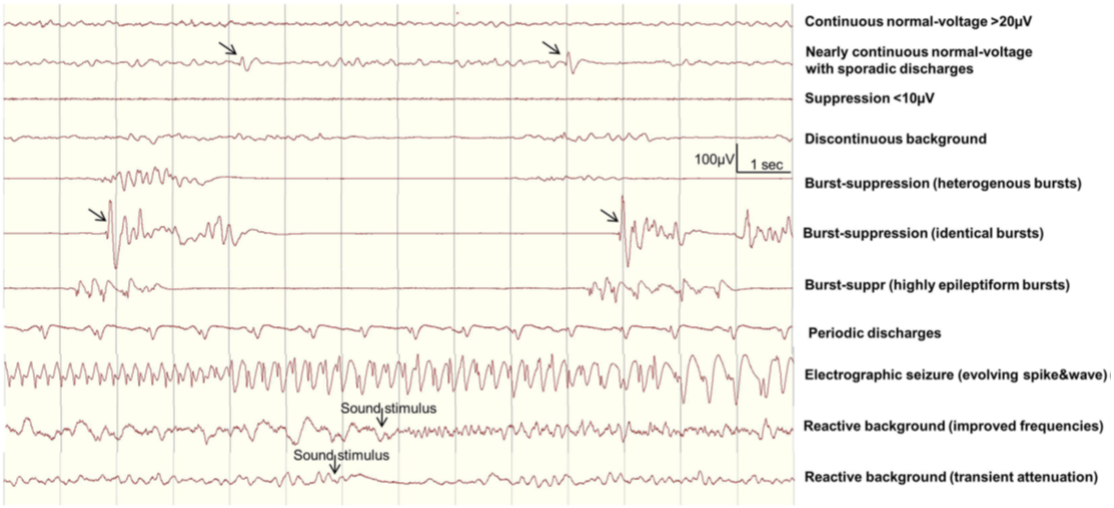

Figure 1.5: EEG patterns (taken from [45]).

In summary, highly malignant patterns include suppressed background with or without superimposed discharges and burst-suppression, especially with identical bursts.

European guidelines on post resuscitation care recommend using EEG for prognostication not earlier than 24 h after ROSC, as confounding factors such as sedation and hypothermia may affect its interpretation [20]. Sedation alters the EEG signal in a dose-dependent manner, reducing its amplitude, frequency and continuity as the dosage increases. However, highly malignant patterns are not induced by standard sedative regimens. Therefore, while sedation should always be considered when analysing EEG, it does not compromise its prognostic value [53]. Similarly, mild hypothermia may influence EEG, but its effects are generally minimal [54]. In addition, some evidence suggests that the prognostic accuracy of EEG can be even higher in the first stages of coma, specifically between 12-24 h after CA, potentially allowing for earlier outcome prediction [55], [56], [57], [58]. These publications highlight that improvements of brain activity within 24 h are crucial for neurological recovery.

1.3.5 Limitations of visual interpretation

Extracting prognostic information from EEG signals is a complex task that cannot be performed by general intensive care staff. In clinical practice, EEG is

analysed visually by a neurologist or a neurophysiologist [59]. This procedure is time-consuming, requires years of expertise and it is prone to both intra- and inter-observer variability [60], [61]. Moreover, a non-negligible number of patients remain in a "grey zone" due to inconclusive signs of either poor or good neurological outcome, making prognostication and clinical decision-making particularly challenging [62]. Another drawback of visual EEG analysis is that it is inherently constrained by the electroencephalographer's ability to interpret a multidimensional time series [63]. Its qualitative nature prevents a comprehensive quantitative assessment, potentially overlooking crucial signal features, such as statistical properties or fine-grained activity patterns, that may carry significant prognostic value [59].

Given these limitations, there is a clear need for novel approaches to improve neurological prognostication. Computational techniques, being fast, automated and objective, represent a promising alternative to visual assessment. Among artificial intelligence methods, deep learning models have rapidly gained increasing interest due to their ability to automatically learn complex features and perform classification directly from raw EEG input.

1.4 Towards objective prognosis: Deep Learning in EEG analysis

1.4.1 Introduction to Deep Learning

In recent years, artificial intelligence has revolutionized the medical field, enabling unprecedented advancements in diagnostics, prognosis and clinical decision-making.

Artificial intelligence comprises any technique that enables computers to perform tasks that usually require human intelligence. Machine learning (ML) is a subset of artificial intelligence that includes all the approaches that allow machines to learn iteratively from data without being explicitly programmed [64].

Among the countless applications of ML, classification is one of the most widely used. Based on the available data, classification algorithms automatically learn the distinctive features of each class (training phase) and use this knowledge to predict the class of previously unseen examples (testing phase). This process is known as supervised learning, as during the training phase, the model is provided with the class labels of each data point, allowing it to learn to map input features to their corresponding output labels. In contrast, during testing, the input of the model is not labelled so that its ability to make predictions can be evaluated.

The training of traditional machine learning algorithms relies on features that must be manually extracted from raw data. These features are designed to represent specific properties of the original data while preserving the information necessary to distinguish between classes. When analysing EEG, features can be derived from

the time, spatial and frequency domains. This process, known as feature extraction or feature engineering, requires a lot of time and expertise. The reliance of ML algorithms performance on experts' knowledge represents one of their greatest strengths and, at the same time, one of the main drawbacks of such approaches [65].

Deep learning (DL), a class of ML techniques, has emerged as a powerful solution to this limitation. Unlike traditional ML algorithms, deep learning models can automatically extract relevant features directly from raw data and perform classification. This approach, known as feature learning, has the major advantage of being data-driven and not constrained by a priori knowledge. For this reason, and due to the complexity of their architectures, DL approaches typically require significantly larger datasets compared to traditional ML methods.

One of the major limitations of DL models, especially in healthcare applications, is their lack of interpretability, which undermines user trust. Their ability to directly process raw data and provide final predictions without clear insight into the decision-making process often makes them a "black box".

Deep learning models are based on Artificial Neural Networks (ANN), whose architecture is inspired by the structure and learning mechanisms of the human brain. These networks consist of layers of interconnected processing units, which represent neurons linked by synapses. Each artificial neuron receives many numerical inputs, each multiplied by a weight and then summed with a bias term. The neuron's output is determined by applying an activation function to its total input, introducing nonlinearity into the model. This mechanism mimics the integration of synaptic signals of variable strength at the cell membrane level and the subsequent generation of an action potential if a certain threshold is reached. The first layer of an ANN receives the data input, the output layer provides the final prediction, while the layers in between, called hidden layers, are responsible for learning a non-linear mapping between input and output.

Convolutional Neural Networks (CNNs) are a class of ANNs specifically designed to process data with a grid-like structure, such as images or time-series, as EEG signal. CNNs differ from ANNs due to the presence of convolutional and pooling layers. Convolutional layers apply filters, also known as kernels, which slide across the width and height of the input data to extract relevant features. The activation layer then introduces non-linearity to the output of the convolutional layer. Finally, the pooling layer reduces dimensionality by summarizing regions of the input using metrics such as the maximum (max pooling) or the average (average pooling). This sequence of layers can be stacked to extract hierarchical features, with the first layers capturing simple patterns and deeper layers gradually learning more complex structures. The final block of a CNN, responsible for classification, consists of layers similar to those in an ANN, known as fully connected (FC) layers, where each neuron is connected to every neuron in the previous layer. During training,

both the filters in the convolutional layers and the weights and biases in the FC layers are optimized to minimize the error between the model's predictions and the true labels (Figure 1.6).

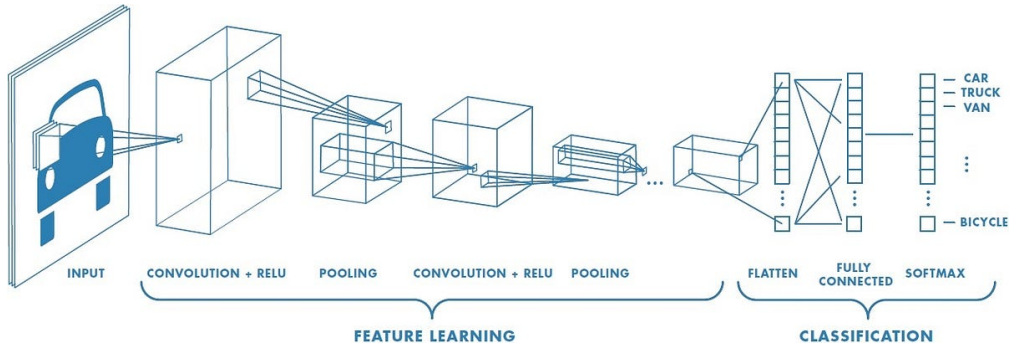

Figure 1.6: CNN architecture (taken from [66]).

1.4.2 State of the art

Several studies in the scientific literature have developed and evaluated deep learning models applied to the EEG recordings of comatose patients following CA, with the aim of predicting their neurological outcome. The vast majority of these studies employed CNNs applied to resting-state EEG, the spontaneous electrical activity of the brain recorded while the patient is not exposed to external stimuli. In every work described below, the outcome was dichotomized based on the CPC score: CPC 1-2 indicated favourable outcome, while CPC 3-5 indicated unfavourable outcome. When evaluating classification performance in this context, it is important to consider that an incorrect prediction of unfavourable outcome may have more dramatic consequences than an overly optimistic one. This is particularly critical when model predictions are used to guide clinical decisions, potentially leading to the early WLST, in cases where recovery may have been possible.

Van Putten et al. [67] designed a CNN with one layer to analyse a monocentric dataset. In particular, the network processed 10-second segments from 5-minute artifact-free EEG recordings at 12 h and 24 h after cardiac arrest. The EEG with electrodes placed according to the 10-20 system, was band-pass filtered between 0.5-35 Hz, re-sampled at 64 Hz and re-referenced based on a longitudinal bipolar montage. Each 10-second segment was classified independently and the predictions of all epochs for a given patient were averaged to obtain the patient's outcome. At 24 h after cardiac arrest, more patient recordings were available ($\sim$400) compared to 12 h ($\sim$280). Outcome prediction was more accurate at 12 h with a sensitivity of 58% at a specificity of 100% for the prediction of unfavourable outcome.

Tjepkema-Cloostermans et al. [68] exploited a Visual Geometry Group (VGG) architecture [69] in a multicentric study. This network consists of a deep CNN with 13 convolutional layers, with a progressively decreasing size of convolutions. The EEG preprocessing was the same as the previous study [67]. Data from two centres (∼660 patients) were used for training and internal validation, while signals from three other centres (∼230 patients) were used to evaluate the network's performance. The classification was performed separately at 12 h and 24 h after cardiac arrest, as well as after combining data from the two time points. A better performance, similar to the previous study [67], was obtained at 12 h. Finally, incorporating data from the 12 and 24 h time points did not significantly modify the model's predictive performance.

A comparative study conducted by Pham et al. [70] compared the performance obtained with the CNN from the previous study [68] with the one achieved by two traditional ML algorithms (random forest and logistic regression) at 12 and 24 h after CA. The dataset was the same as that used by Tjepkema-Cloostermans et al. [68]. For the logistic regression model, two quantitative features were extracted from each 5-minute EEG recording, while for the random forest, nine features were computed. The random forest performance was significantly lower to that of the CNN at each time point and for the prediction of both favourable and unfavourable outcome. Logistic regression yielded comparable or higher performance than the CNN in the first 12 h and lower performance at 24 h. Additionally, this study evaluated the robustness of the three models to superimposed noise added to the EEG signal. The CNN was shown to be less affected by artifacts compared to the other ML techniques.

Building upon CNNs approaches, Jonas et al. [65] used a reduced version of VGG with 6 convolutional layers applied to a monocentric dataset (∼260 patients). For each patient, 5-minute EEG recordings (10–20 system) were downsampled to 50 Hz and were split into overlapping 10-second segments, with a 75% overlap between consecutive segments. Similar to the aforementioned studies, patient-level predictions were obtained by averaging the predictions across individual EEG epochs belonging to that patient. The mean latency from CA was approximately 20 h. Unfavourable outcome was labelled as class 1, while favourable outcome as class 0. The network yielded an accuracy of 83%, a sensitivity of 78% and a specificity of 89%. Although these performances were the highest, they were generally comparable to those achieved on the same dataset by other models, such as the VGG network [68] or the CNN proposed by Van Putten et al. [67]. To better understand which EEG features influenced the model's predictions, Jonas et al. applied gradient-weighted class activation mapping (Grad-CAM), a technique that highlights the parts of the signal that were most informative for the model's decision. The Grad-CAM technique showed that the network relied on similar features to those exploited in clinical EEG visual inspection. Lastly,

the authors enriched the dataset with sleep signals to facilitate the network's ability to recognize favourable patterns and decreasing the number of misclassified unfavourable outcomes. Surprisingly, the opposite results occurred: misclassified favourable outcomes decreased.

Zengh et al. [71] used a large international dataset (∼1000 patients) including continuous EEG recording with electrodes placed according to the 10-20 system from seven different centres. The authors posit that the prognostic value of EEG lies in its temporal evolution following cardiac arrest, rather than in isolated recordings within narrow time windows, as done in previous studies discussed above. To illustrate this, they employed a network called Bidirectional Long Short-Term Memory (BiLSTM), an improved version of a Recurrent Neural Network, which can track and utilize information from both earlier and later points in the input sequence. The raw data were band-pass filtered between 0.5-30 Hz, re-sampled at 100 Hz and re-referenced based on a bipolar montage. Each recording was segmented into consecutive 5-minute intervals and, for each segment, 9 features were extracted. The features extracted at 6-hour recording periods were combined, with the network input consisting of both the features from the current 6-hour epoch and the average features from all preceding 6-hour epochs. The model performance increased with time, achieving the best predicted accuracy at 66 h after cardiac arrest. However, this observation should be interpreted with caution, as it may reflect uneven data availability across time points rather than a true prognostic advantage of this specific window.

A more recent article by Pelentritou et al. [58] employed a CNN inspired by the work of Schirrmeister et al. [72] consisting of three convolutional layers. The authors used a multicentric dataset (∼165 patients) for training and internal validation, along with a portion of the dataset from Zengh et al. [71] for external validation. The authors compared the model performance on the first (9-27 h) and second day (28-56 h) of coma after cardiac arrest, as well as two electrode montage resolutions (62 and 19 electrode configurations). The artifact-free EEG signals, ranging from 8 to 20 minutes in duration and with a sampling frequency of 500 Hz, were segmented into 5-second epochs. The dataset was further augmented through alternating decimation, resulting in epochs with a frequency of 100 Hz. Overall, the network achieved higher performance on the first day of coma for both montages. On the internal validation set, the 62-channel configuration reached the highest accuracy of 94% on day one, compared to 72% on day two. Interestingly, on the second day, the 19-channel montage slightly outperformed the 62-channel configuration, achieving 76% accuracy. The 19-electrode model was also tested on the external validation set, where it achieved an accuracy of 87%. In this work, the use of the Grad-CAM technique revealed that the network's decision-making process is based on evaluations similar to the clinical visual assessment of the EEG. This network demonstrated very high predictive performance even in patients with

uncertain outcome, identified based on inconsistent clinical markers' prediction. Lastly, no significant correlation was observed between predictive performance and sedative levels indicating that differences in the performance between the first and second day of coma were not trivially driven by differences in the clinical management between the two days.

Differently from prior works, an interesting study by Aellen et al. [73] applied a CNN to EEG responses to auditory stimulation during the first day of coma. The model achieved slightly lower performance compared to the aforementioned studies that used resting-state EEG.

Overall, the reviewed publications highlight the promising potential of deep learning models in enhancing the objectivity of comatose patient outcome prediction based on EEG signals. These models achieved comparable or even superior performance relative to earlier computer-based approaches that relied on predefined features, such as functional connectivity [74], [75], [76], power spectra [77], [57] or malignant patterns [78]. Despite encouraging results, further multicentric validation studies are necessary to confirm the reproducibility and robustness of these models across different clinical settings and practices.

The present work was developed in part at the Swiss *Brain-Body and Consciousness* laboratory at the Centre Hospitalier Universitaire Vaudois (CHUV) in Lausanne and it is a natural continuation of the study by Pelentritou et al. [58]. In particular, a CNN, similar to the one proposed by the authors, was applied to a larger and more heterogeneous multicentric dataset, the I-CARE database [79]. The goal of this study was to assess the CNN's performance in predicting post-arrest neurological outcome using EEG recordings acquired between 12 and 24 h after CA, allowing for a more objective and timely prognostic assessment over the first few hours following coma onset. A total of 483 patients were included and the dataset was split into a training set (80%) and an independent test set (20%). Within the training set, a 5-fold cross-validation procedure was used to evaluate the effect of different EEG preprocessing pipelines and CNN parameters on model performance. To identify the optimal EEG preprocessing configuration, signal duration, electrode montage, filtering bandwidth and normalization strategy were tested. Following this exploratory phase, the CNN architecture and its hyperparameters were further optimized using a Bayesian optimization approach. The best performing model following optimization was then retrained on the full training set and evaluated on the held-out test set. In addition to assessing classification performance, interpretability analysis via the Grad-CAM technique was used to investigate the features exploited by the CNN in predicting comatose patient outcome.

Chapter 2

Materials and Methods

2.1 Dataset

2.1.1 I-CARE Database

For the present thesis project, the International Cardiac Arrest Research (I-CARE) consortium database was employed [79], [80]. It represents a real-world dataset including continuous multichannel electroencephalography recordings from 1020 comatose patients following cardiac arrest, along with their neurological outcome. The data were collected by seven different academic hospitals: two from the Netherlands (Medisch Spectrum Twente and Rijnstate Hospital), one from Belgium (Erasme Hospital) and four from the United States (Massachusetts General Hospital, Brigham and Women's Hospital, Beth Israel Deaconess Medical Center and Yale New Haven Medical Center). The two hospitals from the Netherlands are considered part of a single institution, as they are affiliated with the same university.

Neurological outcome was determined prospectively in the two Dutch hospitals by a phone interview at 6 months after ROSC; while, for the remaining medical centres, it was assessed retrospectively through medical chart review at 3-6 months after ROSC. The outcome was measured using the best CPC score. In hospitals without prospective follow-up, patients who achieved a CPC score of 1 or 2 by the time of discharge were considered to have reached their best neurological outcome and no further chart review was performed.

The dataset includes patients older than 15 years old, who underwent cardiac arrest and remained comatose (Glasgow Coma Score ≤ 8) after ROSC. All of them were continuously monitored through EEG after admission to the intensive care unit and for some of them electrocardiogram data was recorded as well. The monitoring usually started within hours after CA and continued for several hours to days depending on the patients' conditions. As a result, the start time and duration of recordings vary across patients. Sedation and analgesia were administered

as needed by clinicians. Sedatives used and typical dose ranges were: propofol (25–80 $\mu g/kg/min$), midazolam (0.1–0.7 $mg/kg/hr$) or fentanyl (25–200 $\mu g/hr$). Neuromuscular blockade during the initiation of TTM was systematically applied in only one centre, while in the remaining hospitals it was administered on an as-needed basis.

EEG data were recorded according to the international 10–20 system, with channel names harmonized across hospitals. When additional electrodes beyond the standard 19 were available, they were also included in the database. For routine clinical procedures, the continuous EEG monitoring was sometimes paused, resulting in occasional gaps in the recordings. No filtering was applied to the signals and their sampling frequency vary between 200 to 2040 Hz. The signals are provided in a unipolar montage, as each channel is labelled with a single electrode. However, the specific reference electrode used is not always documented and can differ between hospitals and patients. Signal quality may deteriorate over time, especially during prolonged recordings.

For each de-identified patient, data are organized into separate files corresponding to consecutive hourly recordings. All data are provided in Waveform Database format, where signal traces are stored in MATLAB files (MAT v4) and metadata in header (`.hea`) files. Metadata include the channels description, signal sampling frequency, power line frequency and start/end times of the recording. For each patient, additional clinical variables are provided in a separate text (`.txt`) file. These include demographic information, cardiac arrest characteristics and neurological outcome. A detailed overview is reported in Table 2.1.

This dataset was used as part of the George B. Moody PhysioNet Challenge 2023 [81]. Of the total 1020 patients originally included, only the data from 607 patients were made publicly available, while the remaining were retained as a hidden test set for the challenge.

2.1.2 Dataset selection

For this work only part of the I-CARE was used. Subjects with at least one recording within 12 to 24 h after CA and duration longer or equal to 30 minutes were considered, resulting in 511 patients. The reason behind choosing latencies in the range of 12 to 24 h is that it has been demonstrated to have high predictive value (subsection 1.3.4), while the second requirement was imposed to ensure a duration of minimum 20 minutes after the application of the preprocessing pipeline. Although the number of patients was reduced from 607 to 511 due to inclusion criteria, the proportion of favourable (FO) and unfavourable outcomes (UO) remained nearly unchanged, with 38% FO and 62% UO in the selected cohort, compared to 37% FO and 63% UO in the original dataset. Hospital C, corresponding to Massachusetts General Hospital, was not represented in the reduced dataset, as none of its patients

Table 2.1: Description of clinical variables provided for each patient in the I-CARE dataset.

Variable	**Description**
Patient	Patient identification number
Hospital	Hospital identifier (A–F)[1]
Age	Patient's age in years
Sex	Biological sex (Male or Female)
ROSC	Time in minutes from CA to ROSC (NaN if not available)
OHCA	`True` = out-of-hospital cardiac arrest `False` = in-hospital cardiac arrest
Shockable rhythm	`True` = initial rhythm was shockable `False` = non-shockable
TTM	Target temperature in °C (33, 36 or NaN if not applied)
Outcome	`Good` = CPC 1–2 `Poor` = CPC 3–5
CPC	Cerebral Performance Category (ordinal scale from 1 to 5)

[1]A=Medisch Spectrum Twente and Rijnstate Hospital; B=Erasme Hospital; C=Massachusetts General Hospital; D=Brigham and Women's Hospital; E=Beth Israel Deaconess Medical Center; F=Yale New Haven Medical Center.

met the inclusion criteria.

To select one recording per patient without introducing bias, recordings were randomly sampled within each latency window while maintaining the original FO/UO proportion from the reduced dataset. The preprocessing strategy, described in detail in the next section, was applied to each selected signal. If, after preprocessing, one or more of the following criteria were not satisfied, a new signal was randomly selected from the remaining available recordings:

- minimum duration of 20 minutes;
- no more than 4 channels interpolated;
- acceptable signal quality (i.e. not excessively noisy or isoelectric).

The automatic random selection was repeated for 3 cycles, after which manual intervention was carried out for final adjustments. In particular, for the remaining patients, among their available recordings, those that met the above inclusion

criteria were selected. Despite efforts to keep the same original outcome distribution, for some latencies the proportions are not perfectly preserved. The final dataset included 483 patients with 39% FO and 61% UO. A summary of the dataset demographic information, stratified by outcome, is reported in Table 2.2. As can be observed, the mean latency for both groups is close to 18 h, which corresponds to the central point of the 12–24 h window and indicates that a balanced distribution was effectively maintained.

Table 2.2: Demographic and clinical characteristics of the final dataset stratified by outcome.

Demographic	**FO (=188)**	**UO (=295)**
Female (%)	51 (27%)	92 (31%)
Age (y)[1] $N_{missing}$[2]	58,2 ± 13,5 1	63,2 ± 15,9 0
Latency (h)[1]	17,96 ± 3,76	17,95 ± 3,74
Shockable (%) $N_{missing}$[2]	135 (72%) 1	104 (35%) 0
OHCA	143 (76%)	211 (72%)
TTM 33°C, 36°C (%) $N_{missing}$[2]	138 (73%), 22 (12%) 28	215 (73%), 28 (9%) 52
Time to ROSC (min)[1] $N_{missing}$[2]	19,4 ± 19,9 131	22,9 ± 19 134
CPC 1 (%) CPC 2 (%) CPC 3 (%) CPC 4 (%) CPC 5 (%)	151 (80%) 37 (20%) - - -	- - 18 (6%) 8 (3%) 269 (91%)
Hospital A Hospital B Hospital C Hospital D Hospital E Hospital F	114 (61%) 23 (12%) 0 (0%) 17 (9%) 11 (6%) 23 (12%)	117 (40%) 71 (24%) 0 (0%) 35 (12%) 38 (13%) 34 (11%)

[1]Mean ± standard deviation.
[2]Number of patients with missing information.

2.1.3 EEG preprocessing

EEG preprocessing was implemented in MATLAB, partially using the FieldTrip toolbox (release 2020-12-05) [82]. First, EEG signals were converted to microvolts (μV). For each channel, the header files contained information on the A/D conversion, such as the gain and the baseline. The gain defines how many digital units correspond to 1 μV, while the baseline indicates the digital value representing 0 μV. So, the formula that allows the conversion from digital to physical units is the following:

$$\text{signal } (\mu V) = \frac{\text{digital signal} - \text{baseline}}{\text{gain}}$$

After conversion, the 19 electrodes from the standard 10-20 system were extracted and ordered to ensure coherence between patients. The resulting EEG data were organized in a matrix with channels as rows and time samples as columns.

Subsequently, the following preprocessing steps were applied to prepare the EEG signals for CNN input:

- **Demeaning** - This step consists of removing the mean value from each EEG channel. The aim is to eliminate the direct current (DC) offset, a constant baseline shift that can result from hardware or recording conditions. By centering the signals around zero, demeaning ensures that subsequent processing steps are not biased by non-physiological baseline variations.
- **Band-pass filtering** - To isolate EEG activity, the raw signal was band-pass filtered between 0.1-40 Hz. Specifically, the band-pass filter was split into a 2^{nd}-order Chebyshev type II high-pass filter (with 30 dB stopband attenuation) followed by a 4^{th}-order Butterworth low-pass filter. The Chebyshev Type II was selected to ensure a steeper cutoff near 0.1 Hz compared to other filter types. Infinite Impulse Response filters were preferred over Finite Impulse Response due to their efficiency and ability to achieve the desired response with lower orders and minimal delay. To prevent phase distortion, all filters were applied in a forward and backward manner (zero-phase filtering). As a result, the effective filter orders were 4^{th} and 8^{th}, respectively.
- **Notch filtering** - A notch filter was applied to attenuate power line frequency, 50 Hz in Europe and 60 Hz in the United States. Although this step was not strictly necessary due to the previous 0.1–40 Hz filtering, it was still applied to ensure the complete removal of power line artifacts.
- **Down sampling** - For uniformity, all EEG signals were down sampled to 200 Hz, the minimum sampling frequency across the dataset.
- **Epoching** - The continuous EEG signals were then segmented into 5-second epochs, resulting in matrices of size 19×1000.

- **Manual channel and trial rejection** - In this step, noisy channels and epochs were removed, improving the global quality of the EEG signal. It was performed using the FieldTrip function `ft_rejectvisual`. In particular, various metrics were computed for each trial and channel and displayed as scatter plots, allowing the user to manually identify and reject outliers (Figure 2.1). The metrics considered in this process were:

  - *Variance* measures the dispersion of the signal amplitudes around the mean. High variance may indicate artifacts, while low variance may suggest flat segments due to disconnected or inactive electrodes.
  - *Inverse variance*, the reciprocal of variance; highlights flat segments or channels.
  - *Minimum, maximum, maximum absolute value and range (max-min)* help detect abnormal and therefore unlikely physiological, peaks and troughs in the signal.
  - *Kurtosis* measures how much the distribution of signal amplitudes deviates from a Gaussian distribution, indicating how likely it is that outliers are present. Extreme values may be associated with artifacts.
  - *Z-value* indicates the distance, in standard deviation, from the mean of the signal distribution. Very high or very low values may indicate artifacts.
  - *Neighbours correlation* identifies channels that have low correlation with each of their neighbours. "Bad" channels usually have inherent noise that is uncorrelated with others.

  It is important to emphasize that rejection decisions did not solely rely on scatter plots, but also on direct visual inspection of the EEG traces. For instance, epileptiform activity or burst-suppression patterns may appear as statistical outliers due to their deviation from background activity, but they are not artifacts and should not be rejected.

- **Bad channel interpolation** - To ensure the same number of input channels for every signal, the channels rejected in the previous step were interpolated. The method used is the spherical spline interpolation by FieldTrip [83]. This approach reconstructs missing channels based on neighbours' activity, considering the head geometry and the electrode placement on a spherical surface.

- **Common average referencing (CAR)** - Since the reference electrodes were not consistently reported across all patients, all EEG signals were re-referenced to the common average. This technique consists of computing the

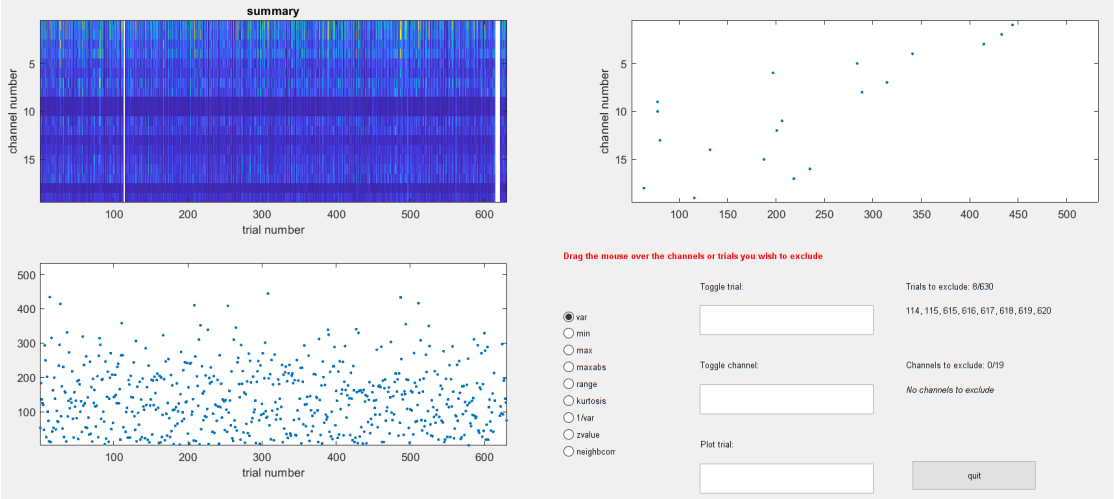

Figure 2.1: FieldTrip interface for the manual trials and channels rejection.

instantaneous mean across all channels and subtracting it from each sample, as follows:

$$CAR(t) = \frac{1}{N}\sum_{c=1}^{N} x_c(t)$$

$$x_c^{CAR}(t) = x_c(t) - CAR(t)$$

where $x_c(t)$ is the signal value of channel c at time t and N is the total number of channels. Additionally, this method improves signal-to-noise ratio by removing the shared noise across channels.

The choice of performing manual channel and trial rejection stems from the complexity and heterogeneity of comatose patients' EEG signals. Typical hypoxic-ischemic brain injury patterns could easily be mistaken for artifacts by automated algorithms. A commonly used technique for artifact removal in EEG is the Independent Component Analysis (ICA), a semiautomatic method that separates statistically independent components, allowing for artifacts removal without eliminating the affected data portions. However, ICA may fail to reliably separate these sources when the number of recording channels is limited, as was the case herein. Additionally, since eye blinks and movement artifacts are largely absent in the EEG of comatose patients, manual preprocessing is considered sufficient to ensure adequate data quality.

The final step of preprocessing was data augmentation, a set of techniques that artificially increase the size and variability of data by generating new samples from existing ones. This is particularly beneficial for deep learning models, which typically require large and diverse datasets to generalize well. In this work, decimation was

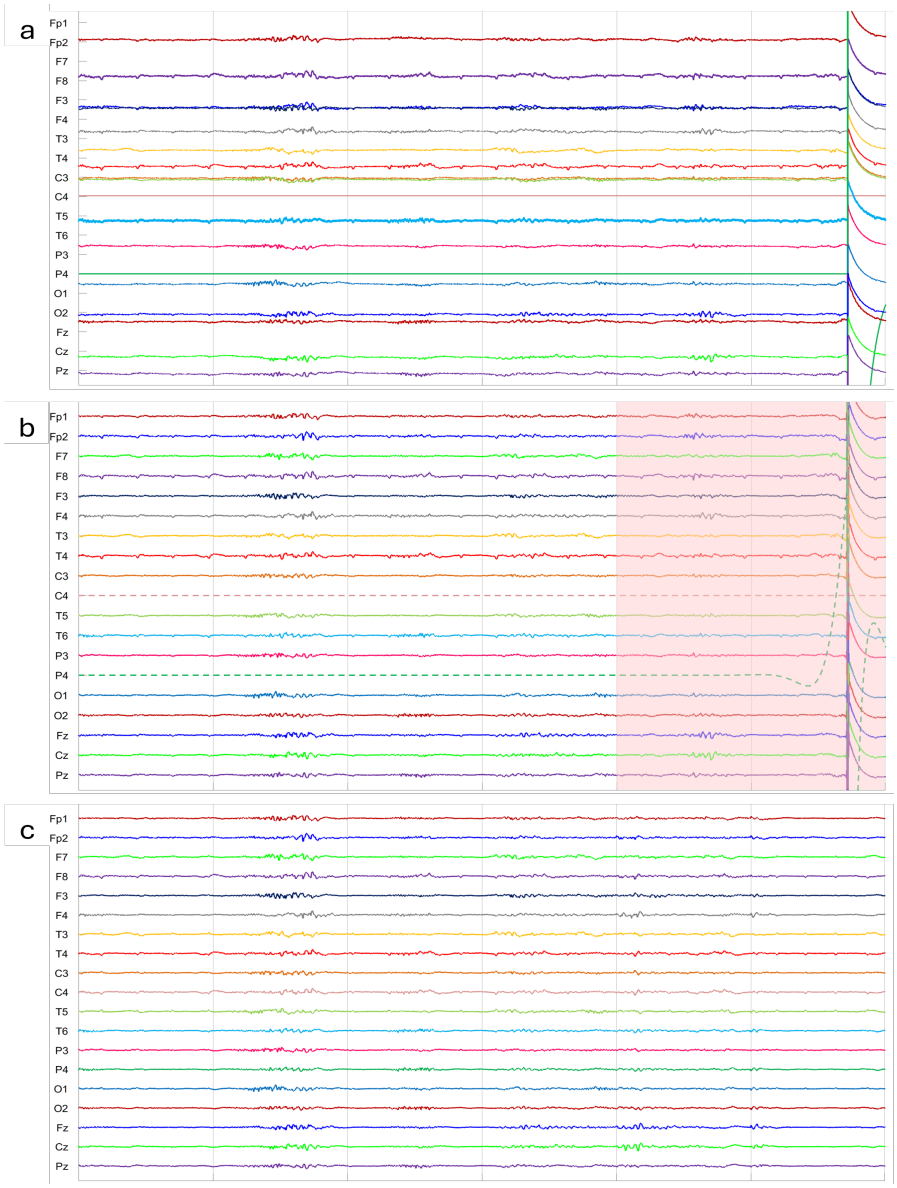

Figure 2.2: A 30-second segment extracted from the EEG of patient 681 (FO), divided into 5-second epochs, is shown at different preprocessing stages. In panel **a**, the raw EEG signal is displayed. Panel **b** shows the signal after demeaning, band-pass and notch filtering and resampling to 200 Hz. In panel **c**, the signal is shown after manual rejection of bad channels and trials, interpolation of removed channels and CAR. In panel **b**, channels C4 and P4 (dashed traces) were removed due to being isoelectric and noisy, respectively. The last two trials (highlighted in red) were excluded due to an artifact, likely of electronic origin, which originated from P4 and propagated to other electrodes. In panel **c**, the application of CAR notably reduced the heartbeat artifact, especially visible in channels F8 and T4 in panel **b**.

employed as a data augmentation strategy. Specifically, each clean 5-second epoch, originally sampled at 200 Hz, was split into two separate 5-second segments at 100 Hz by selecting alternating samples: odd for one segment, even for the other. This procedure effectively doubled the number of dataset examples. Since the signals had already been band-pass filtered between 0.1 and 40 Hz, below the Nyquist frequency of 50 Hz after downsampling, there was no risk of aliasing.

Each 19×500 epoch was associated with the corresponding patient label and used as input to the CNN.

2.2 Convolutional Neural Network

2.2.1 CNN fundamentals

While a brief overview of CNNs has been already provided in the Introduction (subsection 1.4.1), this section offers a more detailed explanation of their internal mechanisms, in order to better contextualize the architecture adopted in this study.

As previously mentioned, what distinguishes CNNs are convolutional and pooling layers. In convolutional layers, filters (or kernels) are small matrices that slide over the input to perform element-wise multiplication followed by summation. The stride defines how many positions the filter moves at each step over the input. When convolution is performed without zero-padding, the output dimensions are reduced; this is referred to as valid convolution. Each convolutional layer can apply a variable number of filters, producing an equal number of feature maps. Each filter is designed to extract different features from the input, allowing the network to learn diverse representations. Next, an activation layer introduces non-linearity into the network, allowing it to learn complex patterns. This is typically followed by a pooling layer, that performs dimensionality reduction by summarizing areas of the feature map superimposed to the moving kernel with their maximum (max pooling) or average (average pooling). These layers can be stacked with varying filter sizes and numbers in order to perform hierarchical feature extraction. At the end of the convolutional and pooling stages, the feature maps are flattened into a one-dimensional vector, which is passed to FC layers, where the actual classification takes place. Final predictions are produced by an activation function, which converts the CNN output into probabilities corresponding to each class.

To train and evaluate the CNN, the dataset is usually divided into three subsets:

- training set;
- validation set;
- test set.

The training process of the model begins with a forward pass, where data from the training set is propagated through the CNN to generate a predicted probability

distribution. At the end, the loss function computes the error between the predicted distribution and the true distribution. The goal of training is to minimize the loss by adjusting the kernels of the convolutional layers and the weights and biases of the FC layers, which are randomly initialized. To determine how to update these parameters, the network computes the gradients of the loss function: the partial derivatives of the loss with respect to the trainable parameters of each layer, which point toward the minimum of the loss function. This is done via backpropagation, which exploits the chain rule to recursively compute the gradients for progressively shallower layers. The parameters are then updated using an optimization algorithm applied to mini-batches of data, with the step size controlled by the learning rate. This whole process is repeated multiple times over the entire dataset, with each complete pass referred to as an epoch.

After each epoch, the model is evaluated on the validation set, a subset of data that is not involved in the training process. This set is used to monitor the model's performance, tune hyperparameters and detect overfitting. Hyperparameters are the set of CNN parameters that are not trainable but must be defined prior to training.

In contrast, the test set is used only after training is complete, to provide an unbiased evaluation of how the model generalizes to unseen data.

2.2.2 CNN architecture

The CNN employed in this work is based on the architecture proposed by Pelentritou et al. [58], which, in turn, was inspired by Schirrmeister et al. [72]. The model is composed of three convolutional blocks followed by two fully connected layers (Figure 2.3). The first convolutional block is specifically designed to handle EEG signals and it is split into two layers:

- The first layer applies temporal convolution, where each kernel performs a 1D convolution across time for each channel independently.
- The second layer performs spatial filtering, where the convolution is applied across all channels simultaneously.

These first two layers aim to extract local temporal and global spatial information from the EEG input, respectively, while the subsequent convolutional blocks are designed to learn both local and global temporal modulations. This architectural choice reflects the nature of EEG: spatially, electrodes capture mixed signals originating from distributed cerebral sources, whereas temporally, brain activity unfolds across multiple scales, with fast local oscillations superimposed on slower global fluctuations [72]. No activation function is applied between these first layers to facilitate the separate learning of temporal and spatial features.

The activation function, applied immediately after the convolutional layers, is the Rectified Linear Unit (ReLU) [84]. It is a simple and wide used function that returns the input value if it is positive and zero otherwise. Then, max pooling is performed with a kernel size of 1x3 and a stride of 3. The number of filters doubles in each convolutional block (25, 50, 100), while the temporal filter size remains the same (1x10), enabling the extraction of increasingly complex features from the signals.

The classification is performed by two fully connected layers. The first layer reduces the feature vector dimensionality and is followed by a ReLU function. The final layer outputs a single value, to which a sigmoid activation function is applied to map the result into a probability score between 0 and 1, where 1 represents favourable outcome while 0 unfavourable outcome. The loss function adopted in this

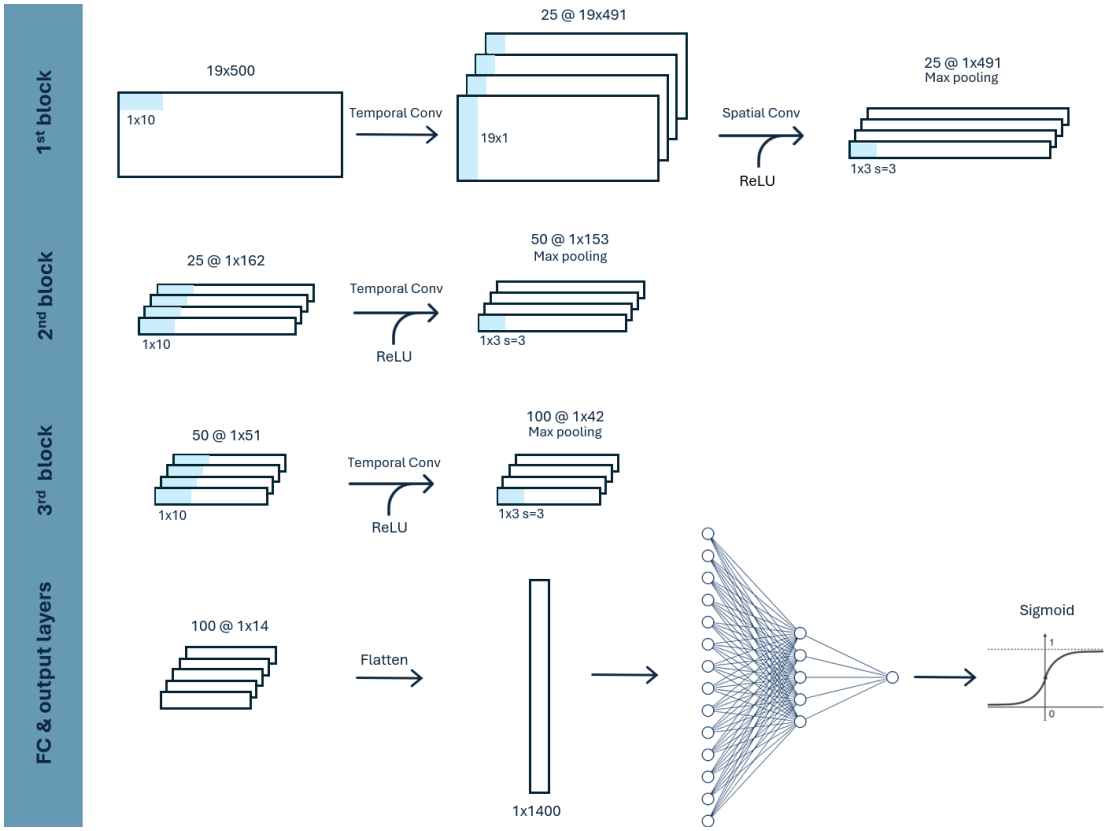

Figure 2.3: Schematic representation of the CNN used in this study.

work is the Binary Cross-Entropy (BCE), a specialized version of the cross-entropy loss tailored for binary classification problems [64]. Cross-entropy quantifies how unexpected the true label, drawn from the data distribution, is with respect to the

predictions made by the distribution modelled by the network. It is given by:

$$\mathcal{L}_{CE} = \sum_{i=1}^{n} \sum_{c=0}^{C} y_{i,c} \log \frac{1}{\hat{y}_{i,c}} = -\sum_{i=1}^{n} \sum_{c=0}^{C} y_{i,c} \log \hat{y}_{i,c}$$

where n is the number of examples in a batch, c refers to the possible class labels of input data, y to the true label and $\hat{y}$ to the model predicted probability.

In the binary case, where $c \in \{0, 1\}$, since the probability of one class is the complement of the other, the cross-entropy loss becomes:

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{n} \left[y \log \hat{y} + (1 - y) \log (1 - \hat{y}) \right]$$

where y is the true label and $\hat{y}$ is the predicted probability of the positive class.

After computing the BCE loss, optimization is carried out using the Adaptive moment estimation (Adam) algorithm [85]. It calculates two statistics of the gradients of the loss function: the first and second moment. The first moment estimates the direction of the gradient via an exponential moving average, which is a weighted average that gives more importance to the current batch's gradient compared to previous ones. Using the average instead of the current value alone allows for a more stable estimation of the direction toward the loss minimum. The second moment is the exponential moving average of the squared gradients. It represents the variability of the gradients and indicates how much the network should move in that direction. This allows the learning rate to be dynamically adjusted: if gradients are large or unstable, the update step is reduced; if they are small or steady, it is increased. The two moments are updated at each training step using the following formulas:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \qquad v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$$

where m is the first moment, v the second, g the gradient and t the number of iterations.

Both moments are bias-corrected (see formulas below) because their moving averages are initialized at zero, causing early updates to be misleadingly small. As the number of iterations increases, the effect of correction diminishes, since $\lim_{t \to \infty} \left(1 - \beta_{1,2}^t\right) = 1$.

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \qquad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

The corrected moment values are then used to update the trainable parameters of the network:

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

where ε is a small value that prevents division by zero and stabilizes the update. The hyperparameters β_1, β_2, ε and the learning rate α must be predefined; in this work they were set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e^{-8}$ and $\alpha = 5e^{-7}$.

In order to enhance training stability and improve generalization, batch normalization and dropout were integrated into the network architecture.

Batch normalization normalizes the output of each convolutional layers before non-linearity, bringing it to zero mean and unit variance across a batch of data. This strategy helps to mitigate the so-called internal covariate shift, the change over time in the distribution of layer outputs caused by the continuous update of network parameters during training [86].

Dropout consists in randomly disabling a fraction of neurons during each training iteration. This prevents the network from becoming too reliant on specific units, thereby improving its ability to generalize and reducing the risk of overfitting. In the CNN adopted in this work, a dropout layer is placed between the two fully connected layers, with a dropout rate set to 0.5.

2.3 Training and evaluation

2.3.1 CNN training and evaluation workflow

The workflow followed for the present thesis is shown in Figure 2.4.

The dataset, composed of 483 patients, was split into two subsets: a training set representing the 80% of the population (386 patients) and a test set comprising the remaining 20% (97 patients). Stratification by outcome was applied in order to avoid bias and ensure that the original UO/FO distribution was preserved across subsets.

To select the best preprocessing configuration and tune the hyperparameters, a 5-fold repeated cross-validation (CV) was employed on the training set. K-fold CV is a common ML strategy in which the training set is divided into k smaller subsets (folds); in each iteration, the model was trained on $k - 1$ folds and validated on the remaining fold. The overall performance of the network was then computed as the average of the performance across all iterations. This procedure allowed for a more robust and stable evaluation of the CNN, reducing its dependency on any specific data splits.

To further enhance the assessment reliability, cross-validation was repeated twice, with different fold splits for each repetition. Although a greater number of repetitions would provide an even more stable performance estimation, two repetitions were considered sufficient to balance computational cost and reliability. Stratification by outcome was applied during all CV splits as well. As a result, a total of 10 distinct training sessions were conducted (5 folds $\times$ 2 repetitions).

After selecting the best signal preprocessing configuration and tuning the CNN hyperparameters, which will be discussed in detail in the following sections, the model was retrained on the entire original training set, without performing any validation. Finally, the performance of the CNN was evaluated on the unseen data of the held-out test set. The following training settings were adopted to ensure both

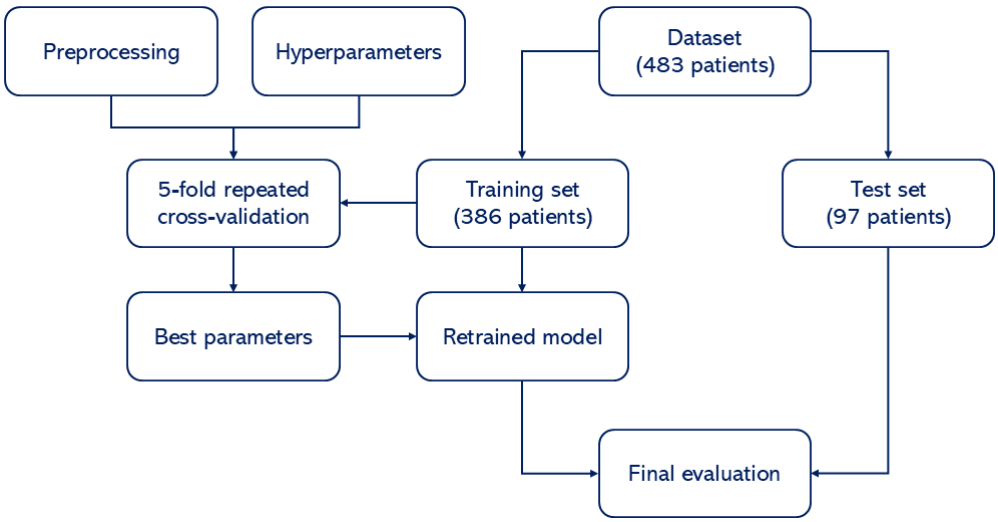

Figure 2.4: Workflow scheme

efficiency and reproducibility. Before each training session in the cross-validation procedure, the model weights were reset to their initial random state to ensure independence among runs.

To avoid overfitting, which occurs when a model fails to generalize because it fits too closely to the training data, early stopping was applied. In particular, after each epoch, the validation loss was monitored and training was stopped if no improvement was observed for 10 consecutive epochs; otherwise, training continued up to a maximum of 100 epochs. The batch size was set to 64.

To prevent that performance changes are attributable to internal sources of randomness within the network, a fixed random seed was set for all Python libraries involved. Additionally, the benchmark mode was disabled and deterministic algorithms were enforced to guarantee that identical inputs always led to identical outputs during training and evaluation.

All CNN training and evaluation procedures were implemented in Python 3.9 using the PyTorch framework (version 1.12). Training was performed on an NVIDIA Tesla V100 GPU with CUDA 11.6, provided by the Legion cluster of HPC@PoliTO (Politecnico di Torino Academic Computing Service) [87].

2.3.2 Evaluation metrics

The network classified each epoch independently, providing a probability value between 0 and 1. To obtain the patient-level prediction, the probabilities assigned to their epochs were averaged. However, these probability scores have to be converted into binary form before computing evaluation metrics. The discriminative threshold for each iteration of the CV was computed through the Receiver Operating Characteristic (ROC) curve derived from the corresponding validation set (Figure 2.5). This graph shows the performance of a binary classifier as the decision thresholds varies: sensitivity (true positive rate) is plotted on the vertical axis and 1-specificity (false positive rate) on the horizontal axis. Considering survival as the positive class, sensitivity (also known as recall) indicates the model's ability to correctly classify patients with a favourable outcome, whereas specificity reflects the ability to correctly identify patients with an unfavourable outcome. An ideal classifier would have a ROC curve reaching the top-left corner of the plot, where both sensitivity and specificity are 100%. The bisector represents the performance of a random classifier. Therefore, a good classifier should have a ROC curve situated between the diagonal and the top-left corner, the closer to the ideal point, the better the performance.

In this study, the cut-off threshold was selected as the one that maximizes the geometric mean (G-mean) of sensitivity and specificity, ensuring a balanced and impartial evaluation of the model's performance:

$$threshold \ = \ max(\sqrt{sensitivity \cdot specificity}\ )$$

An important metric is the area under the ROC curve (AUC), which summarizes the model's ability to distinguish between the two classes: the closer the AUC value is to 1, the higher the discriminative power of the classifier. An AUC value of 0.5 indicates random guessing. It is the only metric computed in this study that does not depend on the threshold definition.

Unlike AUC, all the following evaluation metrics rely on binarized predictions: the probabilities above the threshold were set to 1, while those below to 0. These metrics are derived from the confusion matrix, which summarizes the classification results into four categories: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). TP and TN refer to patients correctly classified as having a favourable and unfavourable outcome, respectively, whereas FP and FN refer to patients incorrectly classified as having a favourable and unfavourable outcome, respectively.

Sensitivity and specificity were computed as follows:

$$Sensitivity \ = \ \frac{TP}{TP \ + \ FN} \qquad Specificity = \ \frac{TN}{TN \ + \ FP}$$

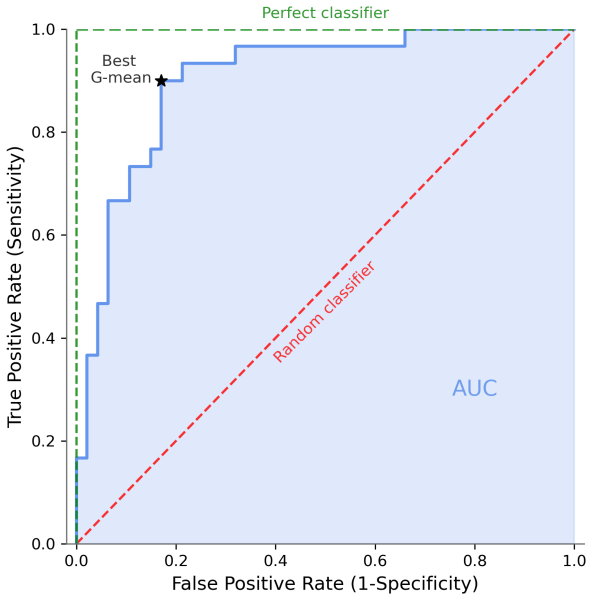

Figure 2.5: Example ROC curve. The dashed red line represents a random classifier, while the green one a perfect classifier. The shaded blue area corresponds to the AUC. The black star marks the point of maximum G-mean, used to select the optimal threshold.

Given the imbalance in the dataset (more unfavourable outcome patients compared to favourable ones), balanced accuracy was preferred over standard accuracy to quantify the model's performance. Indeed, in imbalanced datasets, standard accuracy can be biased toward the majority class. On the other hand, balanced accuracy, defined as the average of sensitivity and specificity, offers a more realistic assessment by giving equal weight to both classes:

$$Balanced\ accuracy\ = \frac{1}{2}\left(\frac{TP}{TP\ + FN}\ +\ \frac{TN}{TN +\ FP}\right)$$

Additionally, two metrics related to sensitivity and specificity were computed: the negative predictive value (NPV) and positive predictive value (PPV), also known as precision. They reflect the reliability of predictions: NPV indicates the likelihood that a patient with a negative test result had UO, while the PPV quantifies the probability that a patient classified as positive had FO. These two metrics are defined as:

$$NPV = \frac{TN}{TN\ +\ FN} \qquad PPV = \frac{TP}{TP\ +\ FP}$$

Finally, the Matthews Correlation Coefficient (MCC) was evaluated. This metric is considered one of the most suitable to measure the quality of classification, particularly for imbalanced datasets, because it takes into account all the elements of the confusion matrix. The MCC ranges from -1 to +1, where +1 indicates perfect

classification, 0 indicates random predictions and -1 indicates inverse classification. It quantifies the alignment between the predictions and the true labels through the following formula:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FN) \cdot (TN + FP)}}$$

All the above-described metrics were computed independently for each CV run on the corresponding validation set. Summary statistics, namely the mean and standard deviation across iterations, were then calculated to provide a robust estimation of the model's performance.

Moreover, validation loss and AUC, together with training loss, were monitored during training epochs to ensure no overfitting occurred.

Model performance was primarily evaluated using AUC, as it provides an overall measure of its discriminative ability regardless of the classification threshold. Sensitivity was considered the second most important metric, given the critical implications of false negatives in the context of comatose patient outcome prediction after CA. Indeed, the misclassification of a patient with a favourable prognosis as having a poor one could potentially lead to premature WLST, resulting in severe ethical and clinical consequences.

2.4 Optimization

2.4.1 Preprocessing configuration selection

The selection of the best preprocessing pipeline configuration was designed as a cascading process: at each step, specific aspects of the preprocessing pipeline were evaluated and, whenever one of the tested options led to improved performance, it was retained for the subsequent steps.

The signal features, for which the impact on the model's performance was analysed, included:

- **Duration** – Three recording lengths were considered: 5, 10 and 20 minutes. Although many patients had more than 20 minutes of clean recording available, the signals were cropped to 20 minutes to ensure the same number of epochs for all patients, as 20 minutes was the minimum recording length across the dataset. This approach helped avoid introducing bias, since more stable and accurate predictions would be expected for patients with a higher number of epochs over which predictive performance is averaged. The 5-minute length was included as it is commonly used in EEG studies for post-cardiac arrest outcome prediction. Finally, the 10-minute segment served as an intermediate duration

to explore performance between shorter and longer recordings. For each patient, a random 5-minute, 10-minute and 20-minute segments was selected from the available recording. This ensured that the model was not biased by any specific portion of the signal and that the analysed recordings were representative of the overall EEG activity. The random selection was different for each duration, and based on the best-performing duration, the same approach was maintained in subsequent steps, ensuring that any differences in performance were not due to different data portions.

- **Data augmentation** - An ablation study was performed by removing data augmentation after selecting the optimal signal duration, in order to assess its individual contribution to model performance. Without augmentation the epochs size increased from 19x500 to 19x1000.

- **EEG montage** - Two different montages were analysed: the unipolar montage with common average referencing and the longitudinal bipolar montage. The latter was included as it is commonly used in pattern analysis following post-cardiac arrest brain injury. The longitudinal bipolar montage is particularly effective in detecting localized abnormalities, as it highlights the differences in electrical activity between adjacent electrodes. To obtain the bipolar montage, the unipolar electrodes with common average referencing were subtracted from each other (Figure 2.6), as this process effectively cancels out the common reference and isolates the differential signal between the electrodes.

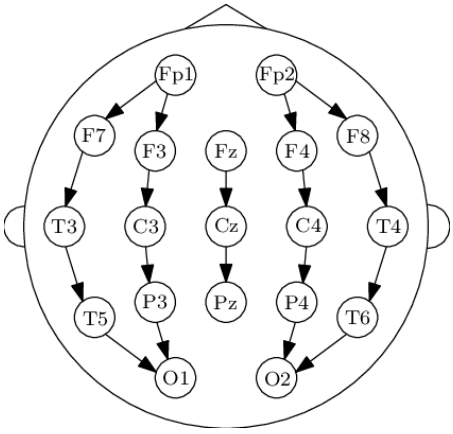

Figure 2.6: Longitudinal bipolar montage. In this configuration the number of resulting channels is 18 (taken from [88]).

- **Filtering bandwidth** - Four bandwidths were included: 0.1-40 Hz, 0.5-40 Hz, 0.1-35 Hz and 0.5-35 Hz. The orders and the types of filters were kept constant, as described in the subsection 2.1.3, only the frequency limits were

changed. Although the tested frequency ranges are relatively close, adjusting the lower and/or upper cut-off frequencies can help reduce low-frequency artifacts or high-frequency noise, potentially impacting the quality of the EEG signal used for classification.

- **Normalization** – CNNs can benefit from dataset normalization, even though it may not always be beneficial depending on the context. Normalizing data changes its distribution, potentially smoothing out differences. In the context of outcome prediction in comatose patients, inter-subject differences are likely informative for the network; therefore, subject-level normalization was discarded. However, reducing variability between channels within the same subject could not remove discriminative information. Two standardization methods were considered: z-score normalization and robust scaling. They were applied to each channel as follows:

$$x_c^{z-norm} = \frac{x_c - \mu_c}{\sigma_c} \qquad x_c^{robust} = \frac{x_c - median(x_c)}{\mathrm{IQR}_c}$$

  where x_c is the channel signal μ_c the mean of the channel, σ_c the standard deviation and IQR_c the interquartile range. Robust scaling, which uses the median and interquartile range, is less sensitive to outliers compared to z-score normalization, which is more effective when applied to Gaussian distributions.

To determine whether the observed performance differences between configurations at each step were statistically significant, a paired Wilcoxon signed-rank test ($\alpha = 0.05$) was applied [89]. The test was performed separately for each evaluation metric by comparing, fold by fold, the results obtained from the two configurations across the repeated 5-fold CV.

2.4.2 Bayesian optimization

Once the best-performing preprocessing pipeline configuration was identified, the subsequent step involved optimizing the CNN using Optuna, a library for automated hyperparameter tuning. Its goal is to identify the combination of parameters that maximizes or minimizes a specific evaluation metric related to model performance. Optuna's default optimizer is Tree-structured Parzen Estimator (TPE), a form of Bayesian optimization. TPE is based on constructing two separate probabilistic models: one for hyperparameter configurations that have shown promising results and another for those that have performed poorly. The initial iterations are used to collect data necessary for estimating these distributions and rely on randomly selected parameter combinations. In later stages, the selection of new configurations is guided by their statistical similarity to those that previously yielded good

performance. This adaptive approach effectively balances the exploitation of promising regions of the search space with the exploration of less-known areas.

In this study the metric that was maximized was the mean AUC, since it reflects the discriminative ability of the model without depending on the threshold.

For computational reasons, in this phase, the number of the 5-fold CV repetitions was decreased from 2 to 1. The number of iterations was set to 50, where the first 10 were used for the probabilistic models' construction.

Given the limited computational resources, the optimization focused exclusively on hyperparameters expected to have the most significant impact on model performance. For the CNN architecture, the optimization involved the initial number of filters, which was subsequently doubled at each deeper convolutional layer, as well as the number of neurons representing the output size of the first FC layer and input size of the second layer. For the training process, the hyperparameters selected for optimization included the learning rate, dropout rate and weight decay.

Weight decay is a regularization technique that helps prevent overfitting and improves generalization by penalizing large weights. In this phase, the AdamW optimizer was used, which applies weight decay directly to the weights at each update step, rather than adding it to the loss function as performed in the more classical weight decay with Adam [90]. This approach avoids modifying the gradients themselves and therefore interfering with adaptive moment estimation, making the regularization effect more consistent and stable. The weights update rule becomes:

$$\theta_{t+1} = \theta_t - \alpha \cdot \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} + \ \lambda \cdot \theta_t \right)$$

where λ is the weight decay coefficient, controlling the strength of the penalty applied to large weights.

It is important to note that, in the initial configuration, Adam optimizer was used without any weight decay ($\lambda = 0$), making it equivalent to AdamW.

Table 2.3: Hyperparameter search space and corresponding sampling methods

Parameter	**Tested values**	**Sampling method**
N. filters	{16, 25, 32}	Categorical
N. neurons	{128, 256, 512, 1024}	Categorical
Learning rate	[$1e^{-7}$ - $1e^{-5}$]	Continuous (log)
Weight decay	[$1e^{-7}$ - $1e^{-4}$]	Continuous (log)
Dropout percentage	[0.2 - 0.6] (step=0.1)	Continuous (linear)

The starting value and the values tested are reported in Table 2.3. This shows that the exploration ranges were generally centred around the initial values to evaluate the impact of both greater and smaller values on CNN performances, with the exception of weight decay, which had previously been fixed to zero. Architectural parameters were defined as categorical, meaning they were sampled from predefined discrete sets. In contrast, training parameters were treated as continuous. Specifically, learning rate and weight decay were sampled from a logarithmic scale, whereas dropout rate was sampled from a linear scale with a fixed discretization step of 0.1. The choice of a logarithmic scale for exploring the learning rate and weight decay allowed for an equal probability of selecting values across different orders of magnitude. This is particularly important due to the model's sensitivity to these hyperparameters: even small changes in them can lead to significant variations in performance.

As outlined above, due to the high computational cost, optimization phase was performed using a 5-fold CV with a single repetition. However, to ensure a fair comparison with the initial configuration the best combination obtained was finally reassessed using a 5-fold CV with 2 repetitions.

To assess whether the performance improvement was statistically significant, a paired Wilcoxon signed-rank test ($\alpha = 0.05$) was applied, as done for the preprocessing optimization [89]. These steps allowed for a more robust and unbiased evaluation of the gains achieved through hyperparameter tuning.

2.5 Final training and testing

The final phase of this study involved retraining the model using the optimized hyperparameter configuration on the entire available training dataset (i.e., all 5 folds), in order to maximize the amount of data used during training before evaluating the model's performance on the unseen test set.

Since no validation set was available in this phase, early stopping could not be applied. Instead, the number of training epochs was fixed to the mean of the epochs at which each run of the repeated CV achieved the best validation performance, using the selected preprocessing pipeline and optimized parameters.

After preprocessing the test data using the optimized preprocessing pipeline, it was fed into the trained CNN. The performance metrics were computed using a decision threshold set as the average of the optimal thresholds obtained across the cross-validation runs.

Finally, to qualitatively assess the model's decision-making process, the Grad-CAM technique was applied on selected epochs from the test set to investigate the EEG portions most relevant in model's prediction.

2.5.1 Grad-CAM

The interpretability of CNN models is of paramount importance to build trust in artificial intelligence, especially in the context of critical care, such as neuroprognostication following cardiac arrest. Understanding which patterns primarily drive the model's decisions helps to demonstrate their consistency with expert encephalographers' knowledge, while also enabling the identification of potential errors or uncovering novel insights that may not emerge from traditional analyses. Unravelling the black-box nature of deep learning models represents a key step toward their future integration into clinical practice.

Techniques such as Grad-CAM represent a powerful means of visually interpreting the decisions made by CNNs [91]. Specifically, Grad-CAM generates a heatmap from the final convolutional layer, highlighting the portions of EEG signals the model mainly relies on to make its predictions.

To compute this image, the feature maps $A^k_{(i,j)}$ resulting from the last convolutional layer are first extracted. Here, k denotes the channel index, and (i, j) the spatial location. The resolution of the heatmap is determined by the size of the feature maps ($Z = H\text{x}W$).

Next, the gradients of the logit (the model's raw output before the sigmoid activation) with respect to the feature maps are computed as $\frac{\partial y}{\partial A^k_{i,j}}$. They represent how strongly the class score is influenced by each activation at each spatial location in the feature maps.

The global average of the gradients is then computed for each feature map, quantifying its importance for the final prediction:

$$\alpha^k = \frac{1}{Z} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y}{\partial A^k_{i,j}}$$

The heatmap is then obtained by computing the weighted sum of the feature maps using the weights α^k, combining their contributions into a single spatial representation:

$$L(i, j) = \sum_k \alpha^k \cdot A^k_{i,j}$$

To retain only the features that contribute positively to the class score and discard those that suppress it, a ReLU function is applied to the heatmap. Since Grad-CAM is class-specific, and the network's output corresponds to the positive class, the heatmap for the negative class can be obtained by applying ReLU to the negated heatmap:

$$L^+(i, j) = \ max(0, \ L(i, j))$$

$$L^-(i, j) = \ max(0, \ -L(i, j))$$

Finally, for visualization purposes, the two heatmaps were normalized with respect to their global maximum and their temporal dimension was upscaled to match the original input length (500), ensuring consistency with the input signal. By comparing the signal epochs with the corresponding heatmaps, it was possible to find out which EEG portions supported classification for both favourable and unfavourable outcomes.

Chapter 3

Results

3.1 Optimization results

Before presenting the results obtained, a general overview of the model training behaviour is provided.

During CV runs, training and validation losses, along with the validation AUC, were monitored across epochs. An example of the learning curves is shown in Figure 3.1. The model typically converged smoothly, with early stopping preventing overfitting by halting training based on the validation loss. These trends were consistent across runs, confirming the stability and effectiveness of the training setup.

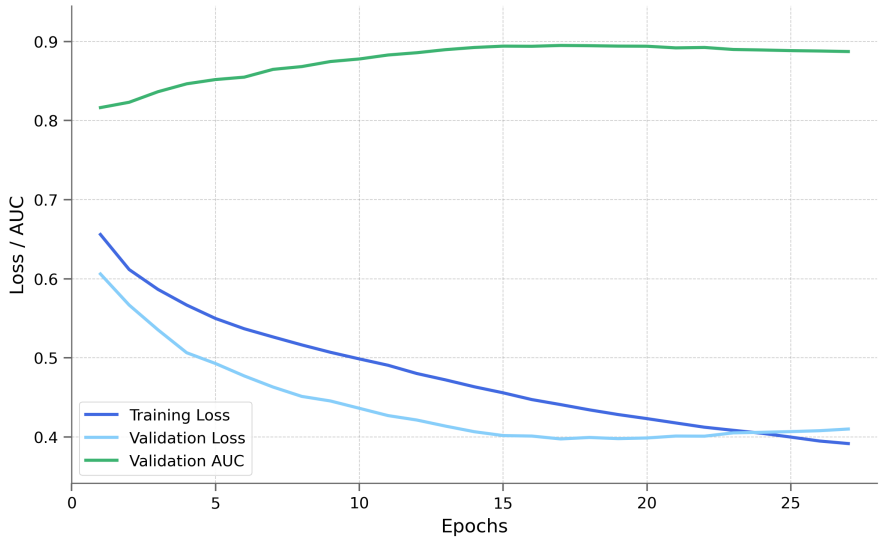

Figure 3.1: Example of training dynamics for one cross-validation run.

3.1.1 Evaluation of EEG preprocessing strategies

As previously described, the process of selecting the optimal preprocessing strategy was conducted in a sequential manner.

Table 3.1: Performance metrics (mean ± standard deviation) and computational time for different EEG recording durations.

Metric	5 min	10 min	20 min
AUC*	0.832 ± 0.055	0.839 ± 0.059	0.844 ± 0.060
MCC	0.517 ± 0.124	0.512 ± 0.129	0.527 ± 0.118
Balanced Accuracy	0.755 ± 0.061	0.756 ± 0.063	0.766 ± 0.058
Sensitivity	0.693 ± 0.102	0.723 ± 0.076	0.760 ± 0.055
Specificity	0.816 ± 0.087	0.788 ± 0.090	0.771 ± 0.092
PPV	0.716 ± 0.098	0.694 ± 0.095	0.689 ± 0.091
NPV	0.810 ± 0.053	0.818 ± 0.048	0.834 ± 0.038
Computational time	1h 50	2h 20	3h 20

*Statistically significant difference in AUC between 5 and 20 minutes ($p - value$=0.049).

The first step focused on assessing the impact of EEG recording duration on model performance. The tested lengths included 5, 10 and 20 minutes. Table 3.1 shows the resulting averages and standard deviations of the evaluation metrics obtained on the validation folds during CV.

Since different random seeds were used to extract epochs for each recording duration, the comparable performance across durations suggests that the model is not sensitive to the specific portions of the EEG used during training. Across different durations sensitivity was consistently lower than specificity, as was PPV compared to NPV. This is potentially a consequence of the dataset imbalance, with more patients having an unfavourable outcome.

The 20-minute duration yielded better performance in nearly all metrics, particularly in AUC (0.844 ± 0.060) and sensitivity (0.760 ± 0.055). Furthermore, the larger number of epochs per subject in the 20-minute condition (480 vs. 240 in 10 *min* and 120 in 5 *min*) contributes to reduced or comparable standard deviation, suggesting slightly more stable and reliable performance estimates.

The paired Wilcoxon signed-rank test revealed a statistically significant difference in AUC between the 5-minute and 20-minute durations ($p-value$=0.049), indicating an improved ability to distinguish between outcome classes with longer recordings. However, no significant differences were found for the threshold-dependent metrics, implying that, once an optimal threshold is applied, the practical benefit of using longer durations may be limited.

The corresponding computational time for each configuration is also reported, showing an expected increase with longer input duration. However, since the additional computational cost remained relatively limited and acceptable, the 20-minute duration was considered a suitable choice for the subsequent steps in the analysis.

In the next step, after the choice of a 20 minute recording duration, the impact of data augmentation on performance was evaluated. The comparison between the CNN results with and without augmentation is presented in Table 3.2. The small difference observed in the evaluation metrics suggests that data augmentation is not strictly necessary in this case, likely due to the dataset being sufficiently large for the model to effectively learn. The paired Wilcoxon signed-rank test confirmed that no statistically significant differences were present between the two configurations. Nevertheless, the model trained with data augmentation achieved a slightly higher AUC (0.844 vs. 0.835), indicating a modest improvement in patient outcome prediction. Interestingly, performance variability remained similar with augmentation, despite its typical association with increased robustness. Given the small numerical advantage in key metrics and the theoretical support for data augmentation in deep learning, this strategy was retained in subsequent analysis steps.

Table 3.2: Performance metrics (mean ± standard deviation) with and without data augmentation.

Metric	**Augmented**	**Non-Augmented**
AUC	0.844 ± 0.060	0.835 ± 0.053
MCC	0.527 ± 0.118	0.523 ± 0.112
Balanced Accuracy	0.766 ± 0.058	0.762 ± 0.054
Sensitivity	0.760 ± 0.055	0.753 ± 0.076
Specificity	0.771 ± 0.092	0.770 ± 0.107
PPV	0.689 ± 0.091	0.690 ± 0.098
NPV	0.834 ± 0.038	0.832 ± 0.043

Next, the unipolar montage with CAR and the longitudinal bipolar montage were compared. As reported in Table 3.3, the unipolar montage yielded slightly better performance across all metrics, with the exception of specificity and PPV. This suggests that slow and spatially distributed cerebral activity, better preserved in the unipolar montage and attenuated in the bipolar configuration, may play an important role in the prognostication of post-anoxic brain injury patient outcome. Of note, the variability across CV iterations was comparable between the two montages. Among the metrics, only AUC showed a statistically significant improvement for the unipolar montage ($p - value$=0.027), reinforcing its potential

advantage in capturing discriminative features.

Table 3.3: Comparison of performance metrics (mean ± standard deviation) between unipolar (CAR) and longitudinal bipolar montages.

Metric	Unipolar (CAR)	Longitudinal Bipolar
AUC*	0.844 ± 0.060	0.826 ± 0.049
MCC	0.527 ± 0.118	0.504 ± 0.119
Balanced Accuracy	0.766 ± 0.058	0.748 ± 0.053
Sensitivity	0.760 ± 0.055	0.710 ± 0.079
Specificity	0.771 ± 0.092	0.787 ± 0.111
PPV	0.689 ± 0.091	0.700 ± 0.118
NPV	0.834 ± 0.038	0.812 ± 0.043

*Statistically significant difference in AUC ($p - value$=0.027).

The impact of the EEG filtering bandwidth on CNN performances was next evaluated. The frequency ranges tested were 0.1–40 Hz (baseline), 0.1–35 Hz, 0.5–40 Hz and 0.5–35 Hz. Results are reported in Table 3.4. While most performance metrics remain relatively similar across the different filtering configurations, AUC and sensitivity showed more variation, offering insight into the possible role of different frequency components. In particular, reducing the low-pass cut-off from 40 Hz to 35 Hz did not substantially affect performance, suggesting that frequencies above 35 Hz may carry limited discriminative information. In contrast, increasing the high-pass cut-off to 0.5 Hz led to a more noticeable decrease in AUC and sensitivity, suggesting a potential contribution of very low-frequency components to the model's predictive performance. This observation is consistent with the previous findings on montage configuration and aligns with existing literature highlighting the role of EEG background activity in neuroprognostication after cardiac arrest. The 0.1–40 Hz configuration showed a statistically significant difference in AUC compared to both 0.5–35 Hz and 0.5–40 Hz, suggesting that the exclusion of very low-frequency components may negatively affect the model's ability to discriminate between classes. However, no statistically significant difference was observed between 0.1–35 Hz and the two 0.5 Hz high-pass configurations. This indicates that the contribution of frequencies below 0.5 Hz, while potentially beneficial, may not be consistently strong across all frequency ranges tested. These results suggest a possible, but not definitive, added value of preserving very low-frequency information.

The configuration with both a higher low-frequency and a lower high-frequency cut-off (0.5–35 Hz) produced mixed results. Overall, the broader frequency range appeared slightly more favourable in terms of class separability and was therefore retained in the final configuration.

Table 3.4: Performance metrics (mean ± standard deviation) across different filtering bandwidths.

Metric	**0.1–40** Hz	**0.1–35** Hz	**0.5–40** Hz	**0.5–35** Hz
AUC*	0.844 ± 0.060	0.837 ± 0.057	0.824 ± 0.060	0.826 ± 0.058
MCC	0.527 ± 0.118	0.513 ± 0.106	0.515 ± 0.106	0.523 ± 0.103
Balanced Acc.	0.766 ± 0.058	0.758 ± 0.056	0.757 ± 0.052	0.760 ± 0.051
Sensitivity	0.760 ± 0.055	0.763 ± 0.118	0.740 ± 0.079	0.730 ± 0.071
Specificity	0.771 ± 0.092	0.753 ± 0.089	0.774 ± 0.112	0.791 ± 0.106
PPV	0.689 ± 0.091	0.671 ± 0.075	0.690 ± 0.093	0.702 ± 0.087
NPV	0.834 ± 0.038	0.839 ± 0.061	0.826 ± 0.039	0.823 ± 0.035

*Statistically significant difference in AUC between 0.1-40 Hz and both 0.5-40 Hz ($p-value$=0.002) and 0.5-35 Hz ($p-value$=0.010).

Finally, the effect of data normalization was explored. Table 3.5 reports the evaluation metrics obtained for the non-normalized dataset, as well as following z-score normalization and robust scaling. Statistical testing did not find any significant difference between the three configurations for any of the evaluated metrics. Nevertheless, both z-score and robust scaling lead to higher specificity and related metrics compared to the non-normalized case, indicating a better ability to correctly identify negative outcomes. Importantly, AUC and sensitivity remained slightly higher without normalization, suggesting that non-normalized input may better preserve features useful for detecting positive cases. Overall, performance variability across CV runs was comparable across the three strategies.

Table 3.5: Performance metrics (mean ± standard deviation) for different normalization strategies.

Metric	**Non-normalized**	**Z-score**	**Robust scaling**
AUC	0.844 ± 0.060	0.836 ± 0.051	0.831 ± 0.039
MCC	0.527 ± 0.118	0.555 ± 0.067	0.550 ± 0.085
Balanced accuracy	0.766 ± 0.058	0.774 ± 0.040	0.771 ± 0.040
Sensitivity	0.760 ± 0.055	0.737 ± 0.121	0.733 ± 0.076
Specificity	0.771 ± 0.092	0.812 ± 0.088	0.809 ± 0.109
PPV	0.689 ± 0.091	0.727 ± 0.071	0.727 ± 0.090
NPV	0.834 ± 0.038	0.836 ± 0.052	0.830 ± 0.034

Despite the marginal differences across configurations, the analysis enabled the identification of a slightly more robust pipeline. The final preprocessing pipeline included 20-minute EEG recordings, in a unipolar montage, filtered between 0.1

and 40 Hz, without normalization and with data augmentation.

For the selected preprocessing pipeline, the ROC curves obtained on each validation set of the CV, along with the mean ROC curve and the corresponding standard deviation are shown in Figure 3.2. The AUC value for each fold is also reported. The mean ROC and AUC and their variability indicate good classification performance and reasonable consistency across different CV runs. While a few folds (e.g., fold 4 and fold 0) yielded slightly lower AUC values, the majority showed a high discrimination performance, suggesting robustness of the selected preprocessing pipeline.

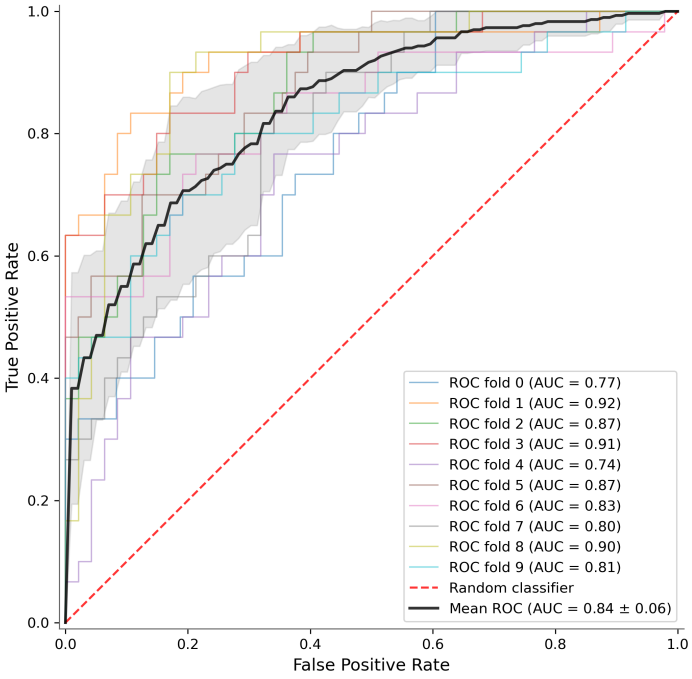

Figure 3.2: Validation ROC curves for each cross-validation fold, computed with the selected preprocessing pipeline. The black line shows the mean ROC, with standard deviation represented by the gray area.

With regards to the optimal thresholds across folds, they showed moderate consistency, with an average of 0.438 ± 0.077 (mean $\pm$ standard deviation). While some variability is expected due to differences in data distribution across folds, the decision boundary remained relatively stable overall.

3.1.2 Evaluation of Bayesian optimization

Bayesian optimization was employed as the final step of hyperparameter tuning. While major improvements were not expected in light of earlier findings, this step

aimed to systematically explore the potential of alternative CNN hyperparameter combinations to improve performance or increase robustness.

The graph in Figure 3.3 shows the history of optimization process across the 50 trials. The objective value, the mean AUC over the 5-fold CV with a single repetition, increased from about 0.83 to almost 0.86.

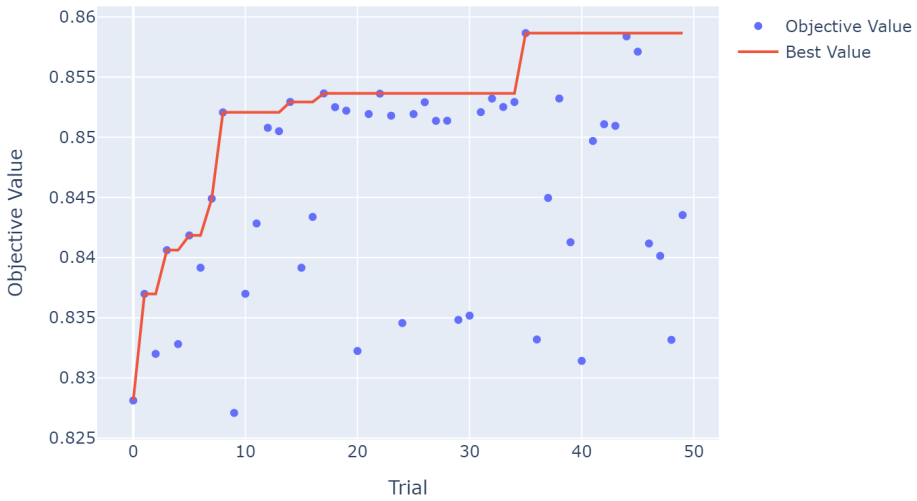

Figure 3.3: CNN hyperparameter optimization history plot

Figure 3.4 shows the distribution of AUC values across the exploration ranges of each tuned CNN hyperparameter during the Bayesian optimization process. This visualization provides insight into the relative influence of each parameter on model performance.

In terms of the number of neurons in the fully connected layer (`n_fc`), a size of 1024 consistently yielded higher performance. Although the flattened input vector has a size of 896, the richness and variability of the extracted features likely benefit from a larger number of neurons, allowing for more effective representation and combination of information.

The optimal initial number of convolutional filters (`n_filter`) was found to be 16. This can be a reasonable choice in lightweight architectures, where a small number of filters is often sufficient to capture low-level patterns, that are progressively enriched by deeper layers. Moreover, using fewer filters reduces the total number of parameters, which helps limit both overfitting and computational cost.

Lower dropout rates (`drop_perc`) are associated with better AUC values. This is likely due to the fact that, in simpler architectures such as the one used here, high rates, causing the deactivation of more neurons, may remove too much information

useful for classification, resulting in a reduction of the model's learning ability.

The learning rate (`lr`) exhibited a non-linear relationship with performance. While values in the 10^{-7} range often resulted in relatively high AUCs, the best-performing trials were achieved with learning rates around 10^{-6}.

For weight decay (`weight_decay`), no clear trend emerged. Performance appeared fairly stable across a wide range of values, suggesting that this regularization parameter had a limited effect on the current model and dataset.

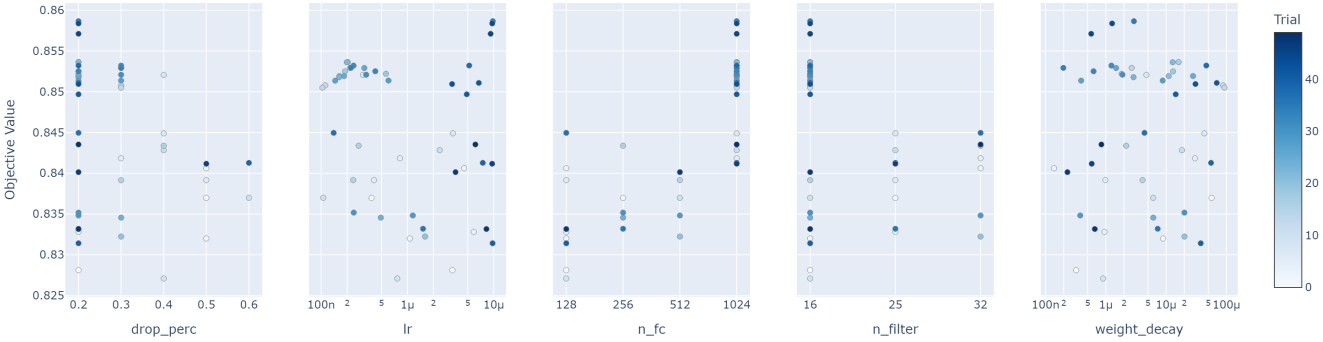

Figure 3.4: Mean AUC distribution across the exploration ranges of each tuned hyperparameter, with each dot representing a trial. The darker the dot, the later the trial in the optimization process.

Given the non-linear effect of the learning rate on performance, and the fact that Bayesian optimization was performed on a single repetition, the best-performing configurations within each of the two most promising orders of magnitude, 10^{-6} and 10^{-7}, were re-evaluated using two 5-fold CV repetitions. In particular, trial 35, the best performing trial from higher learning rates ($AUC_{mean} = 0.858$), and trial 17, the best performing trial from lower learning rates ($AUC_{mean} = 0.853$), were selected for this evaluation.

Aside from learning rate and weight decay, the two configurations share identical hyperparameter values, as reported earlier (`n_fc`=1024, `n_filter`=16 and `drop_perc`=0.2). Table 3.6 summarizes the values of the varying parameters. In trial 17, the regularization effect of weight decay was more pronounced, as its value was higher than in trial 35.

Table 3.6: Parameter differences between trial 35 and trial 17.

Metric	**Trial 35**	**Trial 17**
Learning rate	$9.75e^{-6}$	$2.04e^{-7}$
Weight decay	$2.92e^{-6}$	$1.61e^{-5}$

The evaluation metrics obtained for the two selected configurations are compared with the pre-optimization results in Table 3.7.

Table 3.7: Performance metrics (mean ± standard deviation) before and after optimization.

Metric	Pre-optimization	Trial 35	Trial 17
AUC	0.844 ± 0.060	0.844 ± 0.047	0.844 ± 0.050
MCC	0.527 ± 0.118	0.546 ± 0.061	0.553 ± 0.096
Balanced acc.	0.766 ± 0.058	0.772 ± 0.033	0.773 ± 0.046
Sensitivity	0.760 ± 0.055	0.767 ± 0.104	0.760 ± 0.081
Specificity	0.771 ± 0.092	0.777 ± 0.104	0.786 ± 0.132
PPV	0.689 ± 0.091	0.702 ± 0.073	0.719 ± 0.108
NPV	0.834 ± 0.038	0.847 ± 0.045	0.841 ± 0.033

The tuning process yielded performance metrics similar to the pre-optimization values, as confirmed by statistical testing which found no significant differences, suggesting that the architecture may have reached a performance plateau. However, given the slight improvement observed in the optimized configurations, one of them was selected as the final model.

Trial 17 resulted in higher average performance, especially for MCC, balanced accuracy, PPV and specificity. Variability was in most cases lower compared to in the non-optimized case, but slightly higher than in trial 35.

However, it is worth noting, that the apparent stability of trial 35 may be due to very fast convergence caused by a high learning rate. In each iteration of the CV, validation loss began to increase after just 1 or 2 epochs, triggering early stopping. This dynamic likely prevented the model from fully exploring the loss landscape, increasing the risk of suboptimal convergence and limited generalization. For this reason, trial 17, which underwent a more extended and consistent training process, was selected as the final model.

3.2 Final evaluation on test set

The network with the optimized hyperparameters was retrained with the whole available training dataset (386 patients). The total number of epochs was set to 42, the closest integer to the average epochs at which the best validation performance was observed across the repeated CV folds of the optimized configuration (trial 17).

Finally, the trained model was evaluated on the test set (97 patients). The performance metrics, reported in Table 3.8, were computed placing the cut-off threshold at approximately 0.447, the average threshold obtained across CV runs.

Table 3.8: Performance metrics (mean ± standard deviation) computed on the test set.

Metric	Test set
AUC	0.838
MCC	0.520
Balanced accuracy	0.757
Sensitivity	0.684
Specificity	0.831
PPV	0.722
NPV	0.803

The test set performance was consistent with the average CV metrics of the selected configuration and they fell within the expected variability ranges (Table 3.7, Trial 17). This suggests that the model is stable, did not overfit on training data and is able to generalize effectively on unseen samples.

The corresponding ROC curve is shown in Figure 3.5. Its trajectory is regular without abnormal fluctuations, above the random classification line and rather symmetric, indicating a good balance between sensitivity and specificity. The threshold falls within the steep region (FPR between ≈ 0.1 and 0.3), where a good trade-off between the two metrics can be obtained.

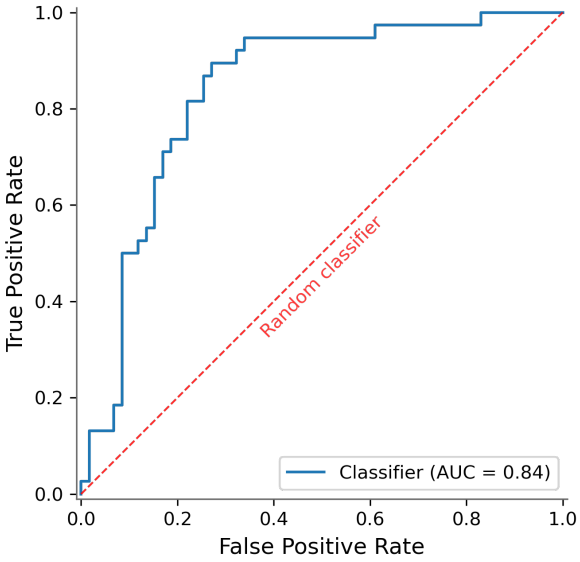

Figure 3.5: Test set ROC curve.

Sensitivity and specificity showed the larger difference compared to the average values across CV folds: sensitivity was slightly lower than the average of 0.760, while specificity was slightly higher than the average of 0.786. The higher specificity compared to sensitivity was potentially driven by the class imbalance in the dataset which made FO patients harder for the network to predict. This is demonstrated in Figure 3.6, which shows the distribution of the predicted probabilities of favourable outcome for patients in the test set, stratified by outcome. FO patients exhibited a more dispersed probability distribution, as indicated by the mean predicted probability being close to the threshold. This contributed to a high number of false negatives, resulting in lower sensitivity and positive predictive value. The majority of correctly classified UO patients had very low predicted probability values, which confirmed the higher confidence of the network in their classification. Nonetheless, false positives, i.e., incorrectly predicted UO patients, were also observed.

Misclassified patients were not only caused by the partial classes overlap around the threshold but also by high-confidence mistakes: patients for whom the CNN assigned a predicted probability of favourable outcome largely different from the ground truth. The underlying reasons for these potentially critical CNN errors will be further explored in the next section through Grad-CAM analysis.

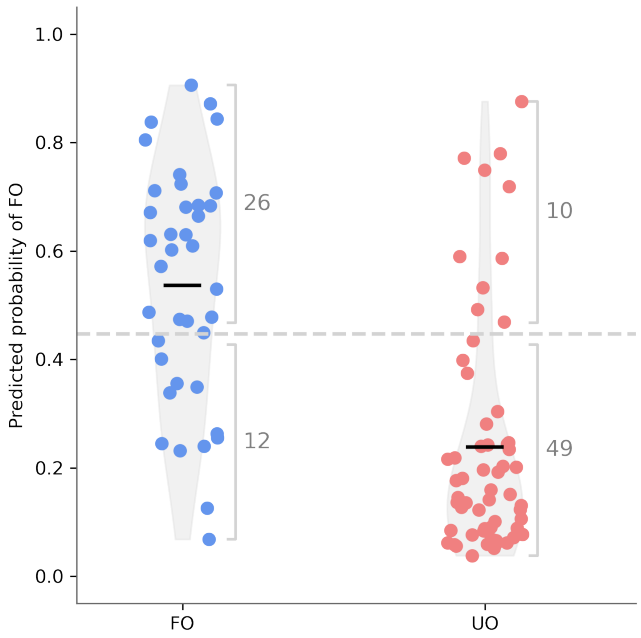

Figure 3.6: Predicted probability distributions for favourable (FO) and unfavourable (UO) outcome. The plot includes the elements of the confusion matrix: TP=26, FN=12, TN=49, FP=10. The black horizontal lines indicate the mean predicted probability for each class, while the dashed line marks the decision threshold.

3.2.1 Grad-CAM analysis

To better understand the classifier's decision-making process and verify whether the features exploited were aligned with neurophysiologists' expertise and not driven by misinterpreted EEG patterns, a qualitative analysis was conducted using the Grad-CAM visualization technique.

This analysis focused on 5-second epochs from a small subset of test set patients. First, correctly classified subjects were examined, followed by patients that the CNN misclassified with high-confidence. The heatmaps shown in this section highlight which EEG portions are discriminative of FO or UO for the CNN. The colour bar reflects the importance of each EEG region in distinguishing between FO and UO.

True negatives

Figure 3.7 shows the Grad-CAM results for patient 1002, correctly classified as having an unfavourable outcome with a predicted FO probability of 0.08. The EEG of this subject was characterized by a burst-suppression background.

The majority of the signal was suppressed with sudden, high-amplitude, jagged bursts. As shown in the Grad-CAM heatmap (a), these bursts were predominantly marked as discriminative of UO. In some cases, the slow deflection between the burst and the suppressed segments was indicative of FO (b).

Figure 3.8 shows the Grad-CAM results for patient number 903, correctly classified as having unfavourable outcome, with a predicted FO probability of 0.09. In this case, the EEG signal exhibited a burst-suppression background with highly epileptiform bursts. The CNN identified the onset of the burst as slightly indicative of FO, likely interpreting it as a possible sign of neural activity after EEG suppression. However, the majority of the burst was strongly associated with UO. The presence of generalized, repeated polyspike-and-wave discharges, followed by an abrupt return to suppression, was indicative of UO for the model. This aligns with expert clinician knowledge since epileptiform activity superimposed on a markedly suppressed background is indicative of severely compromised cerebral function.

Finally, the Grad-CAM heatmaps for patient number 382, correctly predicted as having UO with a predicted FO probability of 0.06, is shown in Figure 3.9. The observed fully suppressed background is strongly discriminative of UO, in agreement with the guidelines upon clinical assessment.

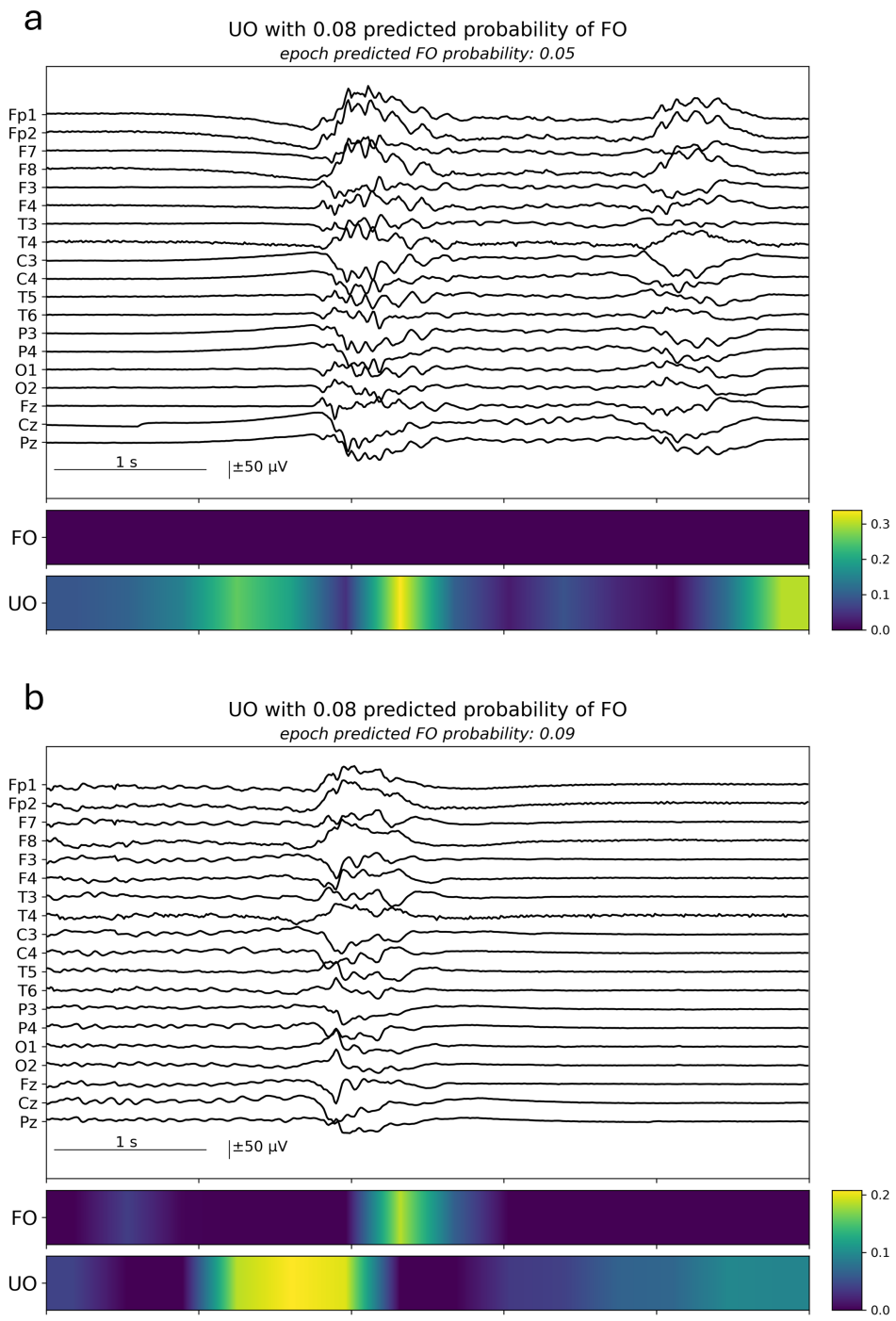

Figure 3.7: Exemplar EEG epochs and corresponding class heatmaps from patient number 1002, a 64-year-old man with an unfavourable outcome (CPC 5), recorded 16 *h* after cardiac arrest.

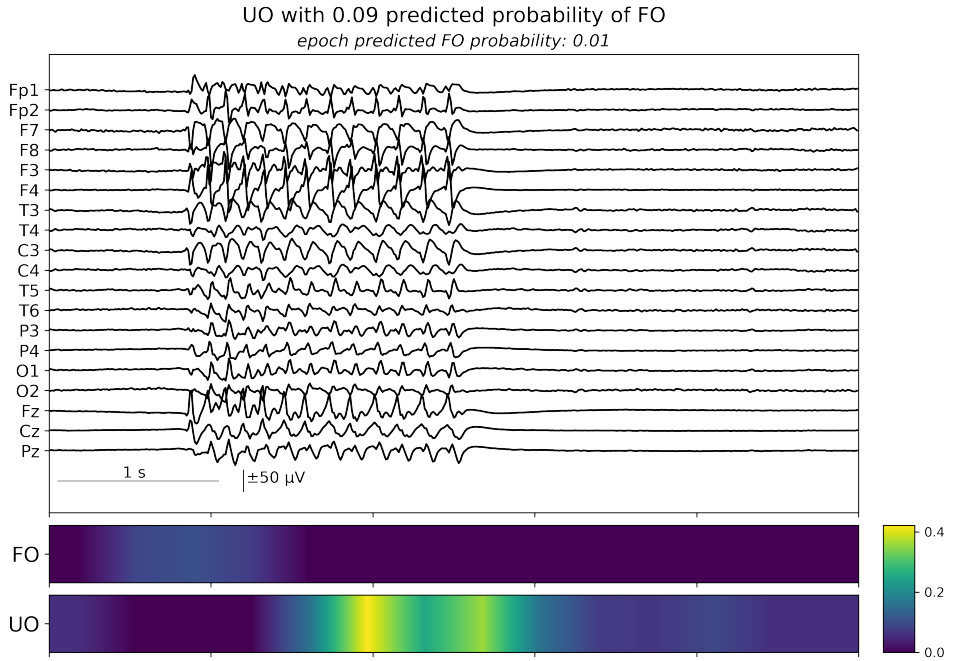

Figure 3.8: Exemplar EEG epoch and corresponding class heatmaps from patient number 903, a 85-year-old woman with an unfavourable outcome (CPC 5), recorded 18 *h* after cardiac arrest.

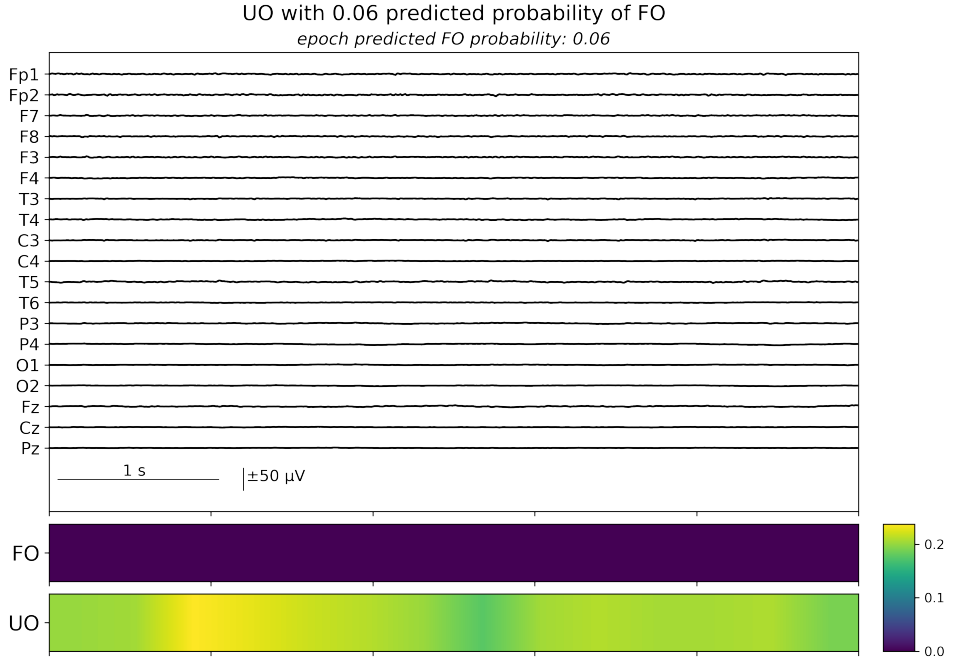

Figure 3.9: Exemplar EEG epoch and corresponding class heatmaps from patient number 382, a 74-year-old woman with an unfavourable outcome (CPC 5), recorded 21 *h* after cardiac arrest.

True positives

EEG benign patterns will be discussed in this section by investigating true positive patients.

Figure 3.10 shows patient number 413, with favourable outcome, correctly classified with a predicted FO probability of 0.91, who presented a continuous theta ($\sim 7\ Hz$) background without superimposed discharges. This was highlighted by the CNN as a strong FO indicator.

A similar observation can be made for favourable outcome patient 584, shown in Figure 3.11, with a predicted FO probability of 0.80. This patient exhibited a continuous, well-modulated alpha background ($\sim 9\ Hz$).

Both examples are consistent with current literature, which regards continuous and organized EEG activity as a marker of FO.

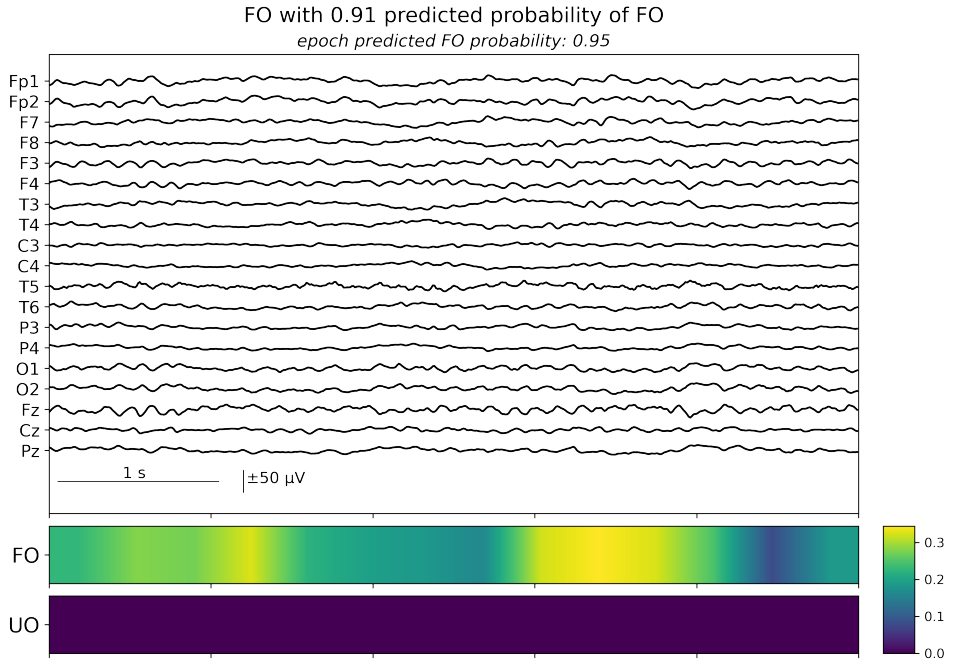

Figure 3.10: Exemplar EEG epoch and corresponding class heatmaps from patient number 413, a 52-year-old woman with a favourable outcome (CPC 1), recorded 17 h after cardiac arrest.

False negatives

False negatives, patients with favourable outcome incorrectly predicted as unfavourable, are of particular interest in the context of the features exploited by the CNN in making a wrong prediction as they may lead to an overly pessimistic prognosis.

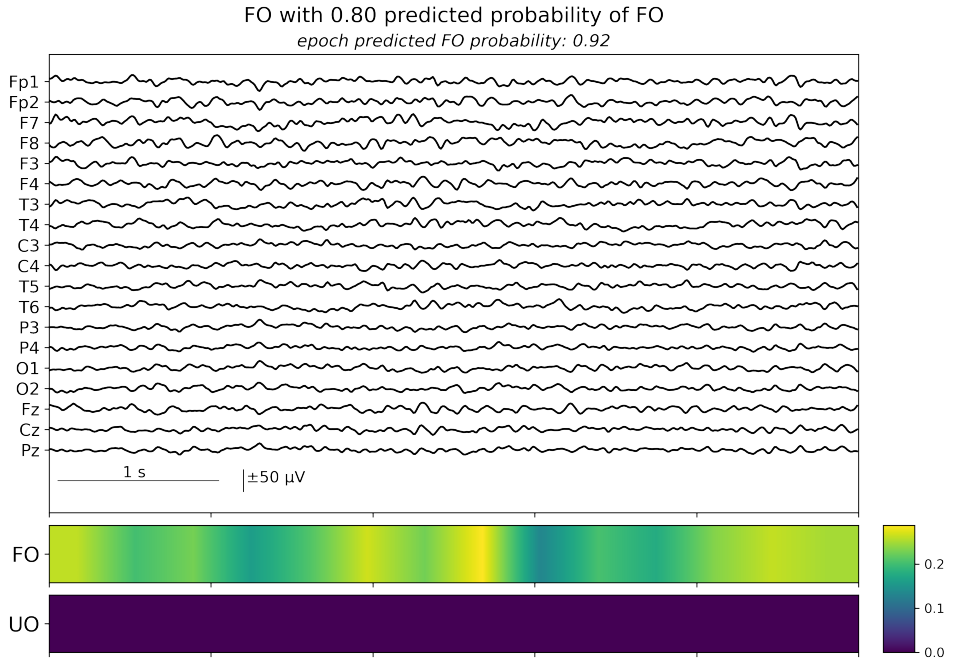

Figure 3.11: Exemplar EEG epoch and corresponding class heatmaps from patient number 584, a 52-year-old man with a favourable outcome (CPC 1), recorded 16 h after cardiac arrest.

Figure 3.12 shows favourable outcome patient 424, who was incorrectly classified as UO with a predicted FO probability of 0.13. The EEG recorded 15 h after CA showed a burst-suppression pattern characterized by prevalent suppression. The bursts, when non-chaotic and characterized by smooth transitions back to suppression, were marked as indicative of FO. However, the suppressed context of these bursts led to an overall unfavourable epoch prediction. Burst-suppression is generally considered a malignant pattern in clinical practise which can nonetheless, improve as time after cardiac arrest passes.

Another example of a false negative prediction is presented in Figure 3.13. Favourable outcome patient 448 was classified as having an unfavourable outcome with a predicted FO probability of 0.07. This can be explained by the isoelectric suppressed background, a highly malignant pattern in clinical EEG assessment, that was also recognized as such by the CNN. The heartbeat artifact is likely the cause of the small regular spikes that appear on the flat signal. The signal from this 72-year-old man may evolve toward benign patterns, that after 15 h from CA are still not visible.

Similarly, other patients who were misclassified as unfavourable outcome patients exhibited suppressed or discontinuous background.

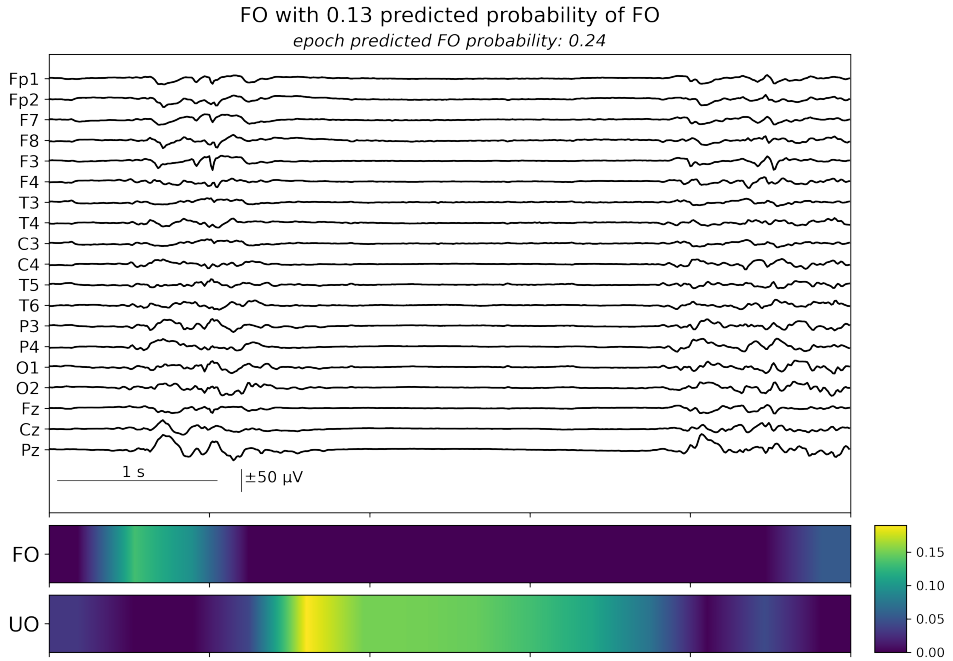

Figure 3.12: Exemplar EEG epoch and corresponding class heatmaps from patient number 424, a 20-year-old man with a favourable outcome (CPC 2), recorded 15 h after cardiac arrest.

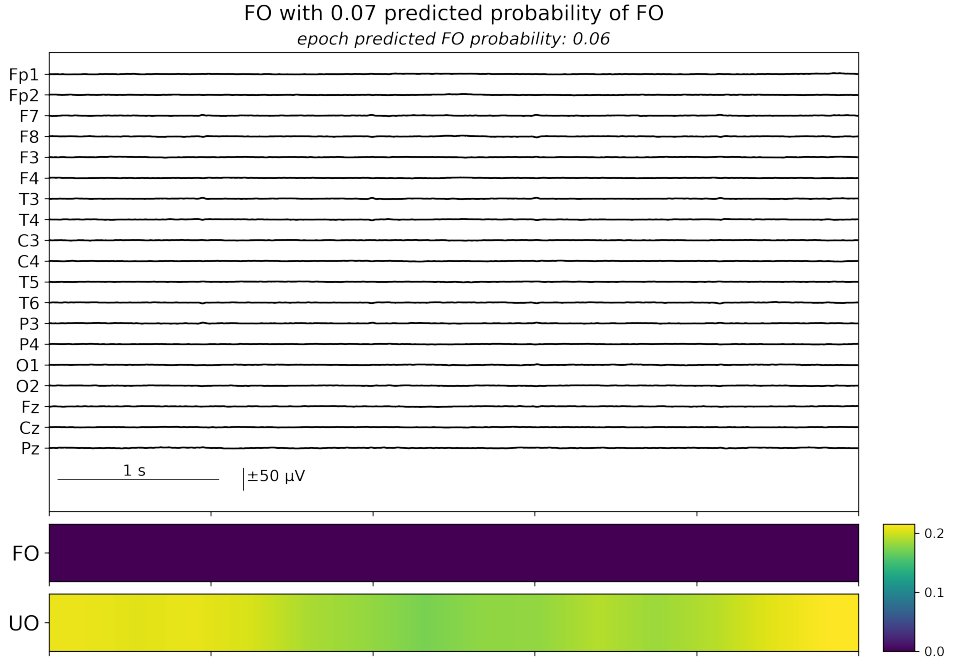

Figure 3.13: Exemplar of EEG epoch and corresponding class heatmaps from patient number 448, a 72-year-old man with a favourable outcome (CPC 1), recorded 15 h after cardiac arrest.

False positives

Of equal importance is the investigation of false positives since such errors in prediction may lead to inappropriate communication with patient relatives or delayed care.

Unfavourable outcome patient 464 is shown in Figure 3.14. The subject, classified as FO with a predicted FO probability of 0.75, showed continuous theta background activity without discharges.

In Figure 3.15, unfavourable patient 991 is presented. The patient was assigned a FO with a predicted FO probability of 0.78. The EEG segment displayed continuous generalised rhythmic theta ($\sim$7 Hz) modulated by delta ($<$ 1 Hz), which is clinically considered as a marker of favourable outcome.

These errors could be due to a transitory recovery, active WLST or perhaps extracerebral causes of death that the CPC score does not explicitly account for. Hence, the neurological prediction of the CNN seems to be coherent with clinical markers and points to the need for case-specific investigations into the network's choices.

Similar to false negative, additional unfavourable outcome patients were incorrectly recognised as having favourable outcome. These patients showed continuous or nearly continuous background without superimposed discharges.

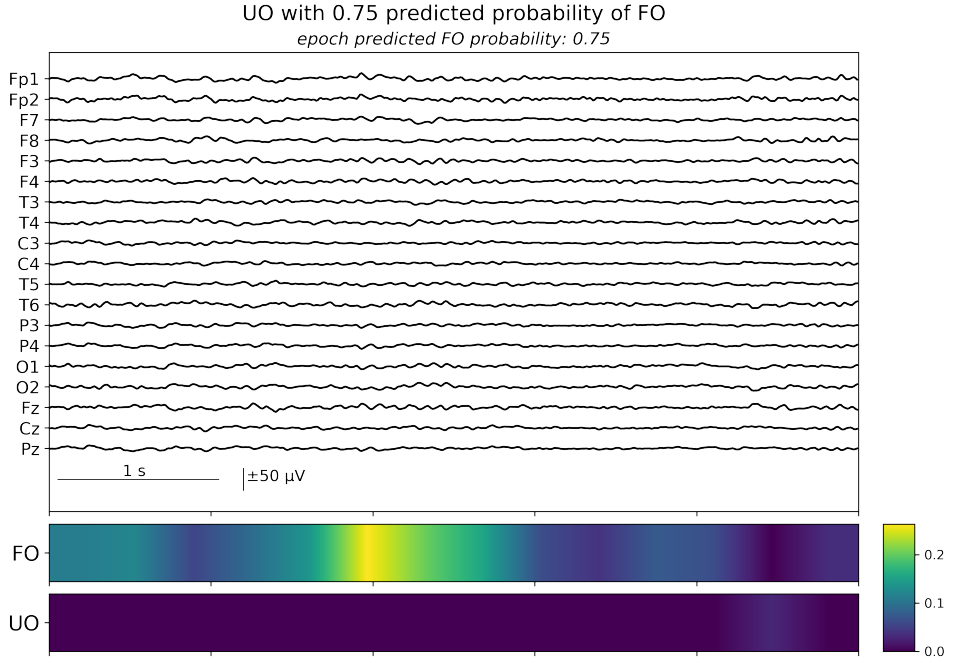

Figure 3.14: Exemplar EEG epoch and corresponding class heatmaps from patient number 464, a 46-year-old woman with an unfavourable outcome (CPC 5), recorded 15 h after cardiac arrest.

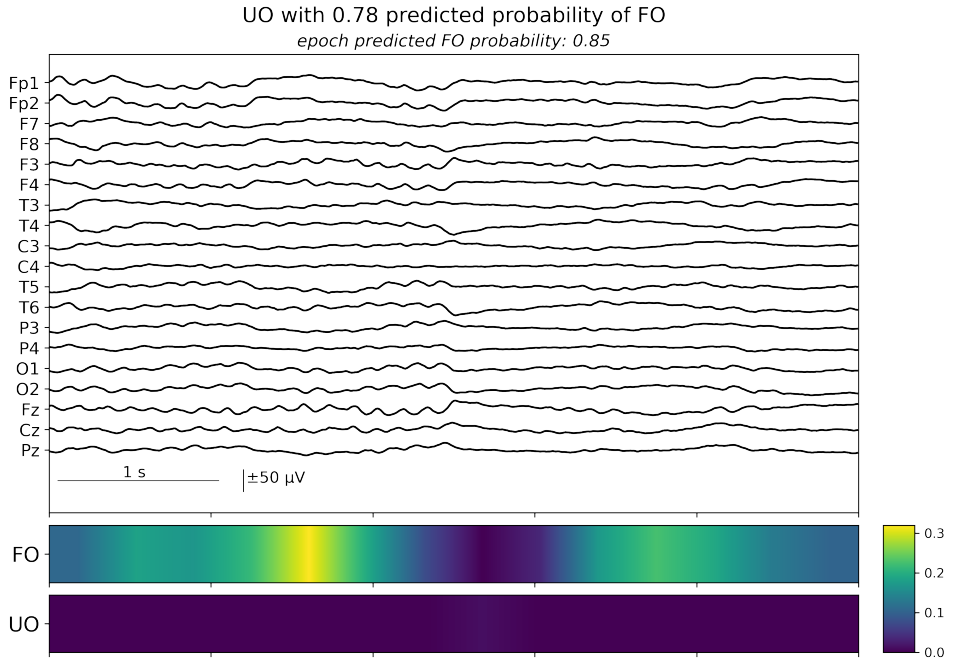

Figure 3.15: Exemplar EEG epoch and corresponding class heatmaps from patient number 991, a 23-year-old man with an unfavourable outcome (CPC 4), recorded 14 h after cardiac arrest.

This analysis suggests that the CNN effectively learned the predictive patterns typically used by clinicians during the visual inspection of the resting-state EEG to assess neurological outcome in comatose patients after cardiac arrest. In addition to classifying suppressed background as unfavourable and continuous background as favourable, Grad-CAM analysis revealed that the network assigned different relevance to bursts depending on their characteristics, suggesting that it may distinguish between subtypes of burst activity rather than treating all bursts uniformly. Consulting a clinical expert could provide further insights into the features that drive these distinctions.

More generally, it seems that mistakes were caused by factors other than the incorrect attribution of prognostic value to EEG patterns. Instead, they might reflect the presence of confounding factors such as extracerebral causes of death, early EEG assessment before potential recovery, WLST, sedation or TTM.

Chapter 4

Discussion

4.1 Summary of the study and main results

This study investigated the use of a DL approach to predict neurological outcome in comatose patients following cardiac arrest, using resting-state EEG recordings. The employed CNN was optimized by selecting the most effective EEG preprocessing strategy and fine-tuning its hyperparameters through Bayesian optimization.

The best-performing preprocessing strategy involved 20-minute EEG segments, band-pass filtered between 0.1–40 Hz, recorded using a unipolar (CAR) montage, with data augmentation applied and without signal normalization.

Bayesian optimization yielded the final set of hyperparameters, which included: learning rate = $2e^{-7}$ (initially $5e^{-7}$), weight decay = $1.6e^{-5}$ (initially 0), number of initial filters = 16 (initially 25), number of neurons = 1024 (initially 512) and dropout rate = 0.2 (initially 0.5). While the architecture remained structurally similar to the original CNN, these adjustments reflect a refinement of its capacity and regularization properties.

During the optimization phase, the model showed stable performance, achieving a mean AUC of 0.844 $\pm$ 0.050 on the validation sets of the 5-fold CV. More importantly, a similar AUC value of 0.838 was obtained on the independent test set, supporting the generalizability of the model.

Beyond predictive accuracy, this work attempted to improve the interpretability of the model's predictions. Grad-CAM, a visualization technique, was applied to highlight the EEG portions with the highest contribution to the network's predictions. The resulting heatmaps showed a correspondence with clinical prognostic EEG markers, suggesting that the model was able to identify physiologically meaningful patterns rather than relying on spurious EEG portions.

Although the different optimization steps did not lead to a substantial performance improvement, the model showed stable and consistent results across

the tested configurations. In critical clinical cases, such as outcome prediction of post-anoxic patients, the ability of a model to maintain consistent performance across varied conditions may be more valuable than marginal gains in accuracy.

The obtained performance is consistent with previous studies using the same multicentric dataset in the 12–24 h window after cardiac arrest, although based on different methodological approaches [71], [92].

Studies with a EEG-based classification design more similar to the present work obtained slightly higher AUC values, ranging from 0.88 to 0.92 [58],[65],[68],[70]. These studies were usually based on smaller datasets, often collected from fewer centres, which may have contributed to more optimistic performance estimates. In contrast, the present work relies on a larger and more heterogeneous dataset, which likely introduces greater variability but also makes the obtained performance more representative of real-world clinical conditions. This highlights the robustness and potential clinical relevance of the proposed approach.

4.2 Clinical integration

The present work adds to the growing literature confirming the potential of DL techniques as a promising complementary tool to visual EEG interpretation for outcome prediction after cardiac arrest. These methods offer fast and objective assessments and can be informative of both favourable and unfavourable outcome, unlike the majority of existing tests [59]. Despite these promising results, several challenges must be addressed before these approaches can be integrated into clinical practice.

First, generalizability remains a key concern and further validation across multiple hospitals is needed to account for variability in EEG acquisition protocols and intensive care clinical practices.

Another challenge is the limited explainability of deep learning algorithms. Methods such as Grad-CAM, as demonstrated herein, can enhance transparency by highlighting the EEG portions that most influenced each prediction. This technique could enable clinical neurophysiologists and neurologists to assess, on a case-by-case basis, whether a given prognosis is based on physiologically meaningful features or non-specific EEG activity. Such transparency could improve trust in automatic methods and facilitate their adoption as clinical decision-support tools.

Additionally, due to differences in outcome distributions between patient populations, defining a fixed classification threshold can be challenging and context-dependent. Therefore, in future clinical applications, the network's output should be treated as a probability rather than a binary decision, allowing clinicians to exploit the full granularity of the model's predictive performance.

Finally, it is worth emphasizing, that neuroprognostication after cardiac arrest

should not solely rely on resting-state EEG evaluation but it should form part of a multimodal assessment including EEG reactivity, clinical examination, biomarkers and imaging [93].

4.3 Limitations and future directions

This study has some limitations which should be discussed. First, the EEG preprocessing step involving the removal of artifacted channels and trials was performed manually. This approach introduces subjectivity and reduces the reproducibility of this work's findings. A full-automated and standardized preprocessing pipeline would instead provide a more robust solution for future applications. Moreover, artifacted channels were interpolated, leading to a partial loss of the original information contained in the signal. Nonetheless, keeping noisy channels in the input data for the CNN would likely result in unsatisfactory model training and ultimately, misclassification of patient outcome.

In addition, only one 20-minute EEG recording was considered per patient, providing a limited look into the brain activity within the 12–24 h period after cardiac arrest. While this choice was motivated by previous literature, it does not exploit the potential richness of information derived from the temporal evolution of brain activity upon the progression of the comatose state. Indeed, recent literature highlights how changes in the resting-state EEG over the course of the injury strongly correlate with clinical outcomes [71]. In light of this, a promising future direction for this work could involve the use of models capable of analysing temporal sequences of EEG data, such as recurrent neural networks (RNNs) or long short-term memory neural networks (LSTMs). These models could be fed with sequential one-hour EEG recordings from the I-CARE database. This approach could potentially offer a more dynamic and physiologically realistic view of the post-anoxic brain recovery process in comatose patients after cardiac arrest.

4.4 Conclusion

In conclusion, this study confirms the great potential of artificial intelligence as an objective and automated tool for predicting neurological outcome in comatose patients after cardiac arrest. The deep learning model proposed herein demonstrated the ability to recognize EEG patterns that are informed by expert knowledge and known to be of prognostic relevance. However, relying on a single "snapshot" of brain activity may not be sufficient to achieve fully reliable predictions. Incorporating the temporal evolution of the EEG could represent a crucial step forward, allowing for the model to capture more complex and physiologically realistic aspects of the patient's condition and in turn improve both performance and clinical utility.

Bibliography

1. Gräsner JT, Wnent J, Herlitz J, Perkins G, Lefering R, Tjelmeland I, Koster R, Masterson S, Rossell-Ortiz F, Maurer H, Böttiger B, Moertl M, Mols P, Alihodžić H, Hadžibegović I, Ioannides M, Truhlář A, Wissenberg M, Salo A, Escutnaire J, and Bossaert L. Survival after out-of-hospital cardiac arrest in Europe - Results of the EuReCa TWO study. Resuscitation 2020; 148:218–26. DOI: 10.1016/j.resuscitation.2019.12.042

2. Zandbergen E, Haan R de, Stoutenbeek C, Koelman J, and Hijdra A. Systematic review of early prediction of poor outcome in anoxic ischaemic coma. The Lancet 1998; 352:1808–12. DOI: 10.1016/S0140-6736(98)04076-8

3. Bayes de Luna A, Coumel P, and Leclercq J. Ambulatory sudden cardiac death: mechanisms of production of fatal arrhythmia on the basis of data from 157 cases. American Heart Journal 1989; 117:151–9. DOI: 10.1016/0002-8703(89)90670-4

4. Myerburg RJ, Kessler KM, Bassett AL, and Castellanos A. A biological approach to sudden cardiac death: Structure, function, and cause. Am. J. Cardiol. 1989; 63:1512–6. DOI: 10.1016/0002-9149(89)90017-9

5. Sinha SK, Moss AJ, and Calkins HG. The etiology of sudden death. *Cardiac Arrest: The Science and Practice of Resuscitation Medicine*. Cambridge University Press, 2007 :229–35. DOI: 10.1017/CBO9780511544828.014

6. Perkins GD, Neumar R, Hsu CH, et al. Improving Outcomes After Post-Cardiac Arrest Brain Injury: A Scientific Statement From the International Liaison Committee on Resuscitation. Circulation 2024. Published online June 27, 2024. DOI: 10.1161/CIR.0000000000001219

7. Semeraro F, Greif R, Böttiger BW, Burkart R, Cimpoesu D, Georgiou M, Yeung J, Lippert F, Lockey AS, Olasveengen TM, Ristagno G, Schlieber J, Schnaubelt S, Scapigliati A, and Monsieurs KG. European Resuscitation Council Guidelines 2021: Systems saving lives. Resuscitation 2021 Apr; 161. Epub 2021 Mar 24:80–97. DOI: 10.1016/j.resuscitation.2021.02.008

8. Buunk G, Hoeven JG van der, and Meinders AE. Cerebral blood flow after cardiac arrest. Netherlands Journal of Medicine 2000; 57:106–12. DOI: 10.1016/s0300-2977(00)00059-0

9. Sandroni C, Cronberg T, and Sekhon M. Brain injury after cardiac arrest: pathophysiology, treatment, and prognosis. Intensive Care Medicine 2021; 47:1393–414. DOI: 10.1007/s00134-021-06548-2

10. Dreier JP, Major S, Foreman B, Winkler MKL, Kang EJ, Milakara D, Lemale CL, DiNapoli V, Hinzman JM, Woitzik J, et al. Terminal spreading depolarization and electrical silence in death of human cerebral cortex. Annals of Neurology 2018; 83:295–310. DOI: 10.1002/ana.25147

11. Leao AA. Further observations on the spreading depression of activity in the cerebral cortex. Journal of Neurophysiology 1947; 10:409–14. DOI: 10.1152/jn.1947.10.6.409

12. Mizoue R, Takeda Y, Sato S, Takata K, and Morimatsu H. Cerebral blood flow threshold is higher for membrane repolarization than for depolarization and is lowered by intraischemic hypothermia in rats. Critical Care Medicine 2015; 43:e350–e355. DOI: 10.1097/CCM.0000000000001095

13. Alevriadou BR, Patel A, Noble M, Ghosh S, Gohil VM, Stathopulos PB, and Madesh M. Molecular nature and physiological role of the mitochondrial calcium uniporter channel. American Journal of Physiology - Cell Physiology 2021; 320:C465–C482. DOI: 10.1152/ajpcell.00502.2020

14. Bonora M, Giorgi C, and Pinton P. Molecular mechanisms and consequences of mitochondrial permeability transition. Nature Reviews Molecular Cell Biology 2022; 23:266–85. DOI: 10.1038/s41580-021-00433-y

15. Cao W, Carney JM, Duchon A, Floyd RA, and Chevion M. Oxygen free radical involvement in ischemia and reperfusion injury to brain. Neuroscience Letters 1988; 88:233–8. DOI: 10.1016/0304-3940(88)90132-2

16. Dux E, Mies G, Hossmann KA, and Siklos L. Calcium in the mitochondria following brief ischemia of gerbil brain. Neuroscience Letters 1987; 78:295–300. DOI: 10.1016/0304-3940(87)90376-4

17. Cirillo C, Brihmat N, Castel-Lacanal E, Le Friec A, Barbieux-Guillot M, Raposo N, Pariente J, Viguier A, Simonetta-Moreau M, Albucher JF, et al. Post-stroke remodeling processes in animal models and humans. Journal of Cerebral Blood Flow Metabolism 2020; 40:3–22. DOI: 10.1177/0271678X19882788

18. Alia C, Spalletti C, Lai S, Panarese A, Lamola G, Bertolucci F, Vallone F, Di Garbo A, Chisari C, Micera S, et al. Neuroplastic changes following brain ischemia and their contribution to stroke recovery: novel approaches in neurorehabilitation. Frontiers in Cellular Neuroscience 2017; 11:76. DOI: 10.3389/fncel.2017.00076

19. Farokhi-Sisakht F, Farhoudi M, Sadigh-Eteghad S, Mahmoudi J, and Mohaddes G. Cognitive rehabilitation improves ischemic stroke induced cognitive impairment: role of growth factors. Journal of Stroke and Cerebrovascular Diseases 2019; 28. DOI: 10.1016/j.jstrokecerebrovasdis.2019.05.023

20. Nolan JP, Sandroni C, Böttiger BW, et al. European Resuscitation Council and European Society of Intensive Care Medicine guidelines 2021: post-resuscitation care. Intensive Care Medicine 2021; 47:369–421. DOI: 10.1007/s00134-021-06368-4

21. Liu Y, Rosenthal RE, Haywood Y, Miljkovic-Lolic M, Vanderhoek JY, and Fiskum G. Normoxic ventilation after cardiac arrest reduces oxidation of brain lipids and improves neurological outcome. Stroke 1998; 29:1679–86. DOI: 10.1161/01.str.29.8.1679

22. Sandroni C, Nolan J, Andersen L, et al. ERC-ESICM guidelines on temperature control after cardiac arrest in adults. Intensive Care Medicine 2022; 48:261–9. DOI: 10.1007/s00134-022-06620-5

23. Lemiale V, Dumas F, Mongardon N, et al. Intensive care unit mortality after cardiac arrest: the relative contribution of shock and brain injury in a large cohort. Intensive Care Medicine 2013; 39:1972–80. DOI: 10.1007/s00134-013-3043-4

24. Laver S, Farrow C, Turner D, and Nolan J. Mode of death after admission to an intensive care unit following cardiac arrest. Intensive Care Medicine 2004; 30:2126–8. DOI: 10.1007/s00134-004-2425-z

25. Sandroni C, D'Arrigo S, Callaway CW, et al. The rate of brain death and organ donation in patients resuscitated from cardiac arrest: a systematic review and meta-analysis. Intensive Care Medicine 2016; 42:1661–71. DOI: 10.1007/s00134-016-4549-3

26. Dragancea I, Rundgren M, Englund E, Friberg H, and Cronberg T. The influence of induced hypothermia and delayed prognostication on the mode of death after cardiac arrest. Resuscitation 2013; 84:337–42. DOI: 10.1016/j.resuscitation.2012.09.015

27. Group BRCTIS. A randomized clinical study of cardiopulmonary-cerebral resuscitation: design, methods, and patient characteristics. American Journal of Emergency Medicine 1986; 4:72–86. DOI: 10.1016/0735-6757(86)90255-X

28. Rankin J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. Scottish Medical Journal 2020; 2:200–15. DOI: 10.1177/00369330 5700200504

29. Rundgren M, Cronberg T, Friberg H, and Isaksson A. Serum neuron specific enolase: impact of storage and measuring method. BMC Research Notes 2014; 7:726. DOI: 10.1186/1756-0500-7-726

30. Karunasekara N, Salib S, and MacDuff A. A good outcome after absence of bilateral N20 SSEPs post-cardiac arrest. Journal of Intensive Care Society 2016 May; 17:168–70. DOI: 10.1177/1751143715616137

31. Benghanem S, Nguyen L, Gavaret M, et al. SSEP N20 and P25 amplitudes predict poor and good neurologic outcomes after cardiac arrest. Annals of Intensive Care 2022; 12:25. DOI: 10.1186/s13613-022-00999-6

32. Scarpino M, Lolli F, Lanzo G, et al. SSEP amplitude accurately predicts both good and poor neurological outcome early after cardiac arrest; a post-hoc analysis of the ProNeCA multicentre study. Resuscitation 2021; 163:162–71. DOI: 10.1016/j.resuscitation.2021.03.028

33. Friberg H, Cronberg T, Dunser M, Duranteau J, Horn J, and Oddo M. Survey on current practices for neurological prognostication after cardiac arrest. Resuscitation 2015; 90:158–62. DOI: 10.1016/j.resuscitation.2015.01. 018

34. Geocadin R, Callaway C, Fink E, et al. Standards for studies of neurological prognostication in comatose survivors of cardiac arrest: a scientific statement from the American Heart Association. Circulation 2019; 140:e517–e542. DOI: 10.1161/CIR.0000000000000702

35. Buzsáki G, Anastassiou C, and Koch C. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. Nature Reviews Neuroscience 2012; 13:407–20. DOI: 10.1038/nrn3241

36. Einevoll G et al. Modelling and analysis of local field potentials for studying the function of cortical circuits. Nature Reviews Neuroscience 2013; 14:770–85. DOI: 10.1038/nrn3599

37. Müller-Putz G. Electroencephalography. *Handbook of Clinical Neurology*. Vol. 168. 2020 :249–62. DOI: 10.1016/B978-0-444-63934-9.00018-4

38. Biasiucci A, Franceschiello B, and Murray M. Electroencephalography. Current Biology 2019; 29:R80–R85. DOI: 10.1016/j.cub.2018.11.052

39. Sutter R and Kaplan P. Electroencephalographic patterns in coma: When things slow down. Epileptologie 2012 Dec; 29:201–9

40. Blinowska K and Durka P. Electroencephalography (EEG). *Wiley Encyclopedia of Biomedical Engineering*. 2006. DOI: 10.1002/9780471740360.ebs

41. St. Louis E, Frey L, Britton J, Frey L, Hopp J, et al. Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants. Chicago: American Epilepsy Society, 2016

42. Teplan M. Fundamentals of EEG measurement. Measurement Science Review 2002; 2:1–11

43. Jasper H. The Ten-Twenty Electrode System of the International Federation. Electroencephalography and Clinical Neurophysiology 1958; 10:371–5

44. Strobbe G. Advanced forward models for EEG source imaging. 2015

45. Sandroni C, Cronberg T, and Hofmeijer J. EEG monitoring after cardiac arrest. Intensive Care Medicine 2022; 48:1439–42. DOI: 10.1007/s00134-022-06697-y

46. Hofmeijer J and Van Putten M. Ischemic cerebral damage: an appraisal of synaptic failure. Stroke 2012; 43:607–15. DOI: 10.1161/STROKEAHA.111.632943

47. Hirsch L, Fong M, Leitinger M, et al. American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2021 Version. Journal of Clinical Neurophysiology 2021; 38:1–29. DOI: 10.1097/WNP.0000000000000806

48. Hofmeijer J, Tjepkema-Cloostermans M, and Van Putten M. Burst-suppression with identical bursts: a distinct EEG pattern with poor outcome in postanoxic coma. Journal of Clinical Neurophysiology 2014; 125:947–54. DOI: 10.1016/j.clinph.2013.10.017

49. Westhall E, Rosen I, Rundgren M, et al. Time to epileptiform activity and EEG background recovery are independent predictors after cardiac arrest. Clinical Neurophysiology 2018; 129:1660–8. DOI: 10.1016/j.clinph.2018.05.016

50. Beretta S and Coppo A. Post-cardiac arrest patients with epileptiform EEG: Better selection for better treatment. Neurology 2020; 94:685–6. DOI: 10.1212/WNL.0000000000009282

51. Muhlhofer W and Szaflarski J. Prognostic Value of EEG in Patients after Cardiac Arrest-An Updated Review. Current Neurology and Neuroscience Reports 2018; 18:16. DOI: 10.1007/s11910-018-0826-6

52. Admiraal MM, Rootselaar AF van, Hofmeijer J, et al. Electroencephalographic reactivity as predictor of neurological outcome in postanoxic coma: A multicenter prospective cohort study. Annals of Neurology 2019; 86:17–27. DOI: 10.1002/ana.25507

53. Ruijter BJ, Van Putten MJAM, Bergh WM van den, Tromp SC, and Hofmeijer J. Propofol does not affect the reliability of early EEG for outcome prediction of comatose patients after cardiac arrest. Clinical Neurophysiology 2019; 130:1263–70. DOI: 10.1016/j.clinph.2019.04.707

54. Hofmeijer J and Van Putten MJ. EEG in postanoxic coma: Prognostic and diagnostic value. Clinical Neurophysiology 2016; 127:2047–55. DOI: 10.1016/j.clinph.2016.02.002

55. Ruijter BJ, Tjepkema-Cloostermans MC, Tromp SC, et al. Early electroencephalography for outcome prediction of postanoxic coma: A prospective cohort study. Annals of Neurology 2019; 86:203–14. DOI: 10.1002/ana.25518

56. Hofmeijer J, Beernink TM, Bosch FH, Beishuizen A, Tjepkema-Cloostermans MC, and Van Putten MJ. Early EEG contributes to multimodal outcome prediction of postanoxic coma. Neurology 2015; 85:137–43. DOI: 10.1212/WNL.0000000000001742

57. Pelentritou A, Ata Nguepnjo Nguissi N, Iten M, et al. The effect of sedation and time after cardiac arrest on coma outcome prognostication based on EEG power spectra. Brain Communications 2023; 5. DOI: 10.1093/braincomms/fcad190

58. Pelentritou A, Gruaz L, Iten M, et al. High density EEG and deep learning improves outcome prediction on the first day of coma after cardiac arrest. medRxiv 2025. Published online January 14, 2025. DOI: 10.1101/2025.01.14.25320516

59. Zubler F and Tzovara A. Deep learning for EEG-based prognostication after cardiac arrest: from current research to future clinical applications. Frontiers in Neurology 2023; 14. DOI: 10.3389/fneur.2023.1183810

60. Westhall E, Rosén I, Rossetti AO, Nielsen N, Ullén S, and Cronberg T. Interrater variability of EEG interpretation in comatose cardiac arrest patients. Clinical Neurophysiology 2015; 126:2397–404. DOI: 10.1016/j.clinph.2015.03.017

61. Benarous L, Gavaret M, Soda Diop M, Tobarias J, De Ghaisne de Bourmont S, Allez C, Bouzana F, Gainnier M, and Trebuchon A. Sources of interrater variability and prognostic value of standardized EEG features in post-anoxic coma after resuscitated cardiac arrest. Clinical Neurophysiology Practice 2019; 4:20–6. DOI: 10.1016/j.cnp.2018.12.001

62. Bongiovanni F, Romagnosi F, Barbella G, et al. Standardized EEG analysis to reduce the uncertainty of outcome prognostication after cardiac arrest. Intensive Care Medicine 2020; 46:963–72. DOI: 10.1007/s00134-019-05921-6

63. Zubler F, Bandarabadi M, Kurmann R, Gast H, and Schindler K. Quantitative EEG in the intensive care unit. Epileptologie 2016; 33:166–72

64. Bishop CM. Pattern Recognition and Machine Learning. New York, NY: Springer, 2006

65. Jonas S, Rossetti AO, Oddo M, Jenni S, Favaro P, and Zubler F. EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. Human Brain Mapping 2019; 40:4606–17. DOI: 10.1002/hbm.24724

66. Saha S. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 Way. Accessed: 2025-04-13. 2018. Available from: https://medium.com/data-science/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

67. Van Putten MJAM, Hofmeijer J, Ruijter BJ, and Tjepkema-Cloostermans MC. Deep learning for outcome prediction of postanoxic coma. *EMBEC & NBC 2017*. Ed. by Eskola H, Väisänen O, Viik J, and Hyttinen J. Vol. 65. IFMBE Proceedings. Singapore: Springer, 2018 :521–4. DOI: 10.1007/978-981-10-5122-7_127

68. Tjepkema-Cloostermans MC, Silva Lourenço C da, Ruijter BJ, and al. et. Outcome Prediction in Postanoxic Coma With Deep Learning. Crit Care Med 2019; 47:1424–32. DOI: 10.1097/CCM.0000000000003854

69. Simonyan K and Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR 2015*. San Diego, CA, 2015 May. DOI: 10.48550/arXiv.1409.1556

70. Pham SDT, Keijzer HM, Ruijter BJ, et al. Outcome Prediction of Postanoxic Coma: A Comparison of Automated Electroencephalography Analysis Methods. Neurocritical Care 2022; 37:248–58. DOI: 10.1007/s12028-022-01449-8

71. Zheng WL, Amorim E, Jing J, et al. Predicting Neurological Outcome From Electroencephalogram Dynamics in Comatose Patients After Cardiac Arrest With Deep Learning. IEEE Transactions on Biomedical Engineering 2022; 69:1813–25. DOI: 10.1109/TBME.2021.3139007

72. Schirrmeister RT, Springenberg JT, Fiederer LDJ, and al. et. Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping 2017; 38:5391–420. DOI: 10.1002/hbm.23730

73. Aellen FM, Alnes SL, Loosli F, Rossetti AO, Zubler F, De Lucia M, and Tzovara A. Auditory stimulation and deep learning predict awakening from coma after cardiac arrest. Brain 2023; 146:778–88. DOI: 10.1093/brain/awac340

74. Kustermann T, Ata Nguepnjo Nguissi N, Pfeiffer C, et al. Brain functional connectivity during the first day of coma reflects long-term outcome. NeuroImage: Clinical 2020; 27:102295. DOI: 10.1016/j.nicl.2020.102295

75. Beudel M, Tjepkema-Cloostermans MC, Boersma JH, and Van Putten MJ. Small-world characteristics of EEG patterns in post-anoxic encephalopathy. Frontiers in Neurology 2014; 5:97. DOI: 10.3389/fneur.2014.00097

76. Zubler F, Steimer A, Kurmann R, et al. EEG synchronization measures are early outcome predictors in comatose patients after cardiac arrest. Clinical Neurophysiology 2017; 128:635–42. DOI: 10.1016/j.clinph.2017.01.020

77. Kustermann T, Nguepnjo Nguissi NA, Pfeiffer C, et al. Electroencephalography-based power spectra allow coma outcome prediction within 24 h of cardiac arrest. Resuscitation 2019; 142:162–7. DOI: 10.1016/j.resuscitation.2019.05.021

78. Müller M, Rossetti AO, Zimmermann R, and al. et. Standardized visual EEG features predict outcome in patients with acute consciousness impairment of various etiologies. Critical Care 2020; 24:680. DOI: 10.1186/s13054-020-03407-2

79. Amorim E, Zheng W, Lee JW, Herman S, Ghassemi M, Sivaraju A, Gaspard N, Hofmeijer J, Van Putten MJAM, Reyna M, Clifford G, and Westover B. I-CARE: International Cardiac Arrest REsearch consortium Database (version 2.1). PhysioNet 2023. DOI: 10.13026/m33r-bj81

80. Amorim E, Zheng W, Ghassemi M, Aghaeeaval M, Kandhare P, Karukonda V, Lee JW, Herman ST, Sivaraju A, Gaspard N, Hofmeijer J, Putten MJAM van, Sameni R, Reyna MA, Clifford GD, and Westover MB. The International Cardiac Arrest Research Consortium Electroencephalography Database. Critical Care Medicine 2023. DOI: 10.1097/CCM.0000000000006074

81. Reyna MA, Amorim E, Sameni S, Weigle J, Elola A, Bahrami Rad A, Seyedi S, Kwon H, Zheng WL, Ghassemi M, Putten MJAM van, Hofmeijer J, Gaspard N, Sivaraju A, Herman S, Lee JW, Westover MB, and Clifford GD. Predicting Neurological Recovery from Coma After Cardiac Arrest: The George B. Moody PhysioNet Challenge 2023. Computing in Cardiology 2023; 50:1–4

82. Oostenveld R, Fries P, Maris E, and Schoffelen JM. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Computational Intelligence and Neuroscience 2011 :156869. DOI: 10.1155/2011/156869

83. Perrin F, Pernier J, Bertrand O, and Echallier JF. Spherical splines for scalp potential and current density mapping. Electroencephalography and Clinical Neurophysiology 1989; 72:184–7. DOI: 10.1016/0013-4694(89)90180-6

84. Nair V and Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning (ICML)*. Haifa, Israel, 2010 Jun :807–14

85. Kingma DP and Ba J. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations (ICLR)*. San Diego, CA, 2015. DOI: 10.48550/arXiv.1412.6980

86. Ioffe S and Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *32nd International Conference on Machine Learning (ICML)*. Lille, France, 2015 :448–56. DOI: 10.48550/arXiv.1502.03167

87. Computing resources provided by hpc@polito, Academic Computing project of the Department of Control and Computer Engineering at Politecnico di Torino. Available from: http://www.hpc.polito.it

88. Borovac A, Runarsson T, Thorvardsson G, and Gudmundsson S. Neonatal seizure detection algorithms: The effect of channel count. Current Directions in Biomedical Engineering 2022; 8:604–7. DOI: 10.1515/cdbme-2022-1154

89. Conover WJ. Practical Nonparametric Statistics. 3rd. John Wiley & Sons, 1999

90. Loshchilov I and Hutter F. Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*. 2019. DOI: 10.48550/arXiv.1711.05101

91. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, and Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 :618–26. DOI: 10.1109/ICCV.2017.74

92. Zhu M, Xu M, Gao M, Yu R, and Bin G. Robust EEG Characteristics for Predicting Neurological Recovery from Coma After Cardiac Arrest. Sensors 2025; 25:2332. DOI: 10.3390/s25072332

93. Rohaut B, Calligaris C, Hermann B, Perez P, Pyatigorskaya N, Galanaud D, Puybasset L, Weiss N, Demeret S, Lejeune FX, Sitt JD, Naccache L, et al. Multimodal assessment improves neuroprognosis performance in clinically unresponsive critical-care patients with brain injury. Nature Medicine 2024; 30:2349–55. DOI: 10.1038/s41591-024-03019-1