
# EEG-Based Brain Age Estimation in Post-Cardiac Arrest Coma

## Domain-Adapted Pipeline Design and Pathological Normative Modelling

Master's Thesis

in Applied Computer Science (M.Sc.)

**Luft Kettenbeil**

Matriculation Number: 21554414

University of Göttingen

May 2026

First Supervisor: Prof. Dr. Julian Kunkel
University of Göttingen

Second Supervisor: Prof. Dr. Nicolai Spicher
Technical University of Denmark

*Ohne Kontext stellen sich die falschen Fragen.*

*Lacking context, we ask amiss.*

# Acknowledgements

*I would like to thank my supervisors,* ***Prof. Dr. Julian Kunkel*** *and* ***Prof. Dr. Nicolai Spicher****, for their guidance and support.*

*I am furthermore grateful to* ***Dr. Narges Lux*** *for making me think aloud, to* ***Steffen Klemer*** *for believing in me, and to* ***Toni Hosp*** *for always making me smile.*

This research was supported by computational resources and infrastructure provided by the **KISSKI project** *(BMBF grant 01 IS 22 093 A-E).*

## Abstract

**Motivation and Objective.** Machine-learning neuroprognostic models in the ICU context are typically trained on outcome data. However, withdrawal-of-life-sustaining-therapy remains the primary cause of in-hospital death for comatose post-cardiac arrest patients following resuscitation. Training models on such data naturally reinforces any false-positive clinical decision. To circumvent this ethical confound, this thesis proposes an outcome-independent biomarker: the *Coma-Normative Brain Age Gap* (CN-BAG). By estimating chronological age from EEG signals relative to the post-cardiac arrest population, the resulting CN-BAG establishes a training target that is causally upstream of withdrawal decisions.

**Methodology.** A domain-adapted, non-destructive preprocessing architecture is developed and tested across five declarative profiles on the multi-centre I-CARE dataset ($N_{usable} = 604$). A two-phase evaluation strategy with strict patient-level cross-validation is employed: feature-based machine learning ensembles capturing spectral, non-linear complexity, and burst-suppression features, followed by end-to-end deep learning (CNN-BiLSTM-Attention) operating on continuous EEG.

**Results.** Across all profiles and architectures, age-prediction models collapse into constant-mean predictors ($R^2 \approx 0$, MAE $\approx 12.2$ years). A sex classification positive control fails to reach the 0.65 feasibility threshold (Phase 1 best: BAcc $= 0.577$, 95 % CI [0.50, 0.63], permutation $p = 0.001$; Phase 2 deep learning: BAcc $= 0.500$, degenerate), suggesting that the collapse is biological rather than algorithmic. However, a sensitivity analysis excluding CPC 5 patients (in whom WLST confounding is most acute) reveals a nominally significant association between the bias-corrected CN-BAG and neurological recovery in the survivor cohort (CPC 1–4, $N = 162$; $p = 0.009$, hypothesis-generating), while the primary full-cohort analysis remains non-significant ($p = 0.058$).

**Conclusion.** These rigorously validated null results formally establish *state-trait eclipse* as the operative constraint in acute coma: the neurophysiological variance introduced by acute hypoxic-ischaemic injury and pharmacological sedation effectively overwrites the premorbid biological identity. Critically, this collapse is not a deep-learning artefact — Phase 1 classical ML, which is less dependent on training-set volume than deep learning, converges on an identical constant-mean prediction, implicating biological rather than purely statistical limits. Ultimately, this thesis demonstrates that while technical noise cleaning in the ICU is highly achievable, the comatose brain's chronological identity appears to be physiologically inaccessible under the acute post-cardiac-arrest conditions studied.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis investigates whether the ageing signal in electroencephalographic (EEG) recordings survives the severe neurological insult of post-cardiac-arrest coma — and whether such a metric may hold the potential to serve as an objective, decision-independent biomarker in clinical neuroprognostication.

This chapter provides the primary motivation for taking this approach, presents the central epistemological challenges, and introduces the Research Objectives. It then summarises the major contributions and provides an outline for the remainder of this thesis.

## 1.1 Clinical Context and Motivation

Machine-learning neuroprognostication in post-cardiac-arrest coma is currently trapped in a lethal feedback loop. The primary in-hospital cause of death for these patients is not irreversible neurological injury, but the clinical decision to withdraw life-sustaining therapy (WLST) following a perceived poor prognosis [1–4]. In published cohorts, up to 52 % of all initially resuscitated patients who die before hospital discharge do so following WLST [1, 3].

When prognostic algorithms are trained directly on survival outcomes (e.g., the Cerebral Performance Category, CPC), they conflate the neuropathological severity of the condition with the clinical decision to withdraw care. Deployed prospectively, such algorithms risk predicting clinicians' withdrawal decisions rather than biological recovery potential. A false-positive poor-outcome prediction is uniquely dangerous in this setting: if it triggers WLST, the patient dies and the prediction appears correct—a self-fulfilling prophecy that artificially inflates the model's apparent accuracy [5].

These two interpretations — accurate neuroprognostication and accurate prediction of clinicians' withdrawal decisions — are formally indistinguishable from observed outcome labels alone. A model that achieves high prognostic accuracy on such data cannot, by construction, establish which of the two it has learned. Furthermore, the standard evaluation metric conceals this feedback loop: false positives among patients who underwent WLST are invisible because the dataset records them as poor outcome, making the measured false-positive rate a lower bound on the true biological error rate.

For a formal Bayesian decomposition of how WLST artificially inflates prediction sensitivity, see Appendix D.

## 1.2 The Coma-Normative Brain Age Gap

To dismantle this circularity, the field must abandon human-annotated clinical outcomes as training labels. This thesis targets an outcome-independent biomarker: chronological age. Chronological age is a biologically prior ground truth, logically prior to the clinical decision-making process, and entirely unaffected by the cardiac arrest or its intensive care management.

This thesis proposes the *Coma-Normative Brain Age Gap (CN-BAG)*. Rather than transferring models trained on healthy individuals—which systematically misinterpret profound pathological slowing as youth (the Rejuvenation Artefact)—this framework trains age regressors exclusively within the post-cardiac-arrest population. The reference shifts from *how does this brain compare to a healthy brain?*—a question that cannot be meaningfully answered in coma—to *how does this brain compare to other equally critically ill brains?* The CN-BAG estimates chronological age from comatose EEG relative exclusively to age-matched comatose peers. The resulting prediction residual quantifies an individual's neurophysiological deviation from the expected pathological trajectory, establishing a proxy for hypoxic-ischaemic severity that is structurally immune to WLST label contamination.

Furthermore, CN-BAG is understood not as a pure measure of biological ageing, but as a *compound biomarker*. A coma-normative age prediction intrinsically weaves together the patient's premorbid neurophysiological constitution (trait), the acute severity of the hypoxic-ischaemic injury (state), and the specific pharmacodynamics of administered sedatives like propofol. This thesis explicitly owns this compound nature (Section 6.4) rather than treating it as a confound to be corrected.

## 1.3 The Problem: Transfer Barriers in ICU EEG

Applying healthy-norm brain age models to comatose ICU EEG confronts severe transfer barriers. This domain shift introduces scientific, methodological, and practical challenges; however, the most pressing barrier is epistemological. Table 1.1 summarises the key challenges at a glance.

**Table 1.1:** Healthy-domain transfer fails at independent levels for post-cardiac-arrest comatose EEG.

| Level | Mechanism | Reference |
|---|---|---|
| Signal domain | Burst-suppression, isoelectric silence, absent alpha rhythms — pathological patterns qualitatively unlike any healthy EEG state | [6] |
| Rejuvenation Artefact | Pathological slowing mimics young-brain spectral signatures, producing systematic age underestimation unrelated to biology | [6] |
| Aperiodic confound | The 1/f slope shifts with injury severity, sedation, and therapeutic hypothermia rather than with chronological age | [7–9] |
| Cardiac contamination | ECG volume conduction dominates the aperiodic slope in ICU EEG; removing cardiac artefact *reduces* apparent age-prediction accuracy in uncleaned models | [9] |
| No healthy control | No normothermic, unmedicated, alert EEG from a healthy brain under ICU recording conditions is obtainable | — |

Beyond these transfer barriers lies a deeper epistemological problem. Brain age is a single-value estimator anchored in typicality: it quantifies how a given brain compares to the expected neurophysiological profile for its chronological age. In a severely pathological state, however, this construct becomes difficult to interpret. The central question is whether brain age and brain health can be meaningfully decomposed — whether state and trait remain separable once acute hypoxic-ischaemic injury has compressed the EEG into global slow-wave pathology.

Even if coma-normative age prediction proves feasible, demonstrating clinical utility

requires a further step: establishing not merely statistical association but a causally interpretable link between the CN-BAG and neurological recovery potential. Feasibility is a necessary but not sufficient condition for clinical deployment.

## 1.4 Research Objectives

Despite those challenges, bypassing the WLST confounder remains a necessity. This thesis therefore operates on the following primary research objectives to establish CN-BAG and to determine its feasibility, validity, and significance:

- **RO-1 (Coma-Normative Brain Age Gap Analysis):** Evaluate whether chronological age can be reliably predicted from comatose ICU EEG, and determine if the resulting CN-BAG predicts neurological outcome independently of age and established clinical predictors. This includes binary sex classification as an *information-theoretic positive control*.
- **RO-2 (Domain-Adapted Preprocessing):** Design, implement, and validate a preprocessing and feature extraction framework purpose-built for comatose ICU EEG, utilising hardware-anchored, state-invariant thresholds to preserve pathological variance.
- **RO-3 (Preprocessing Impact Assessment):** Empirically quantify how different preprocessing strategies (ranging from raw signal to aggressive artefact removal) impact downstream predictive accuracy and feature validity.

For the complete, formal mapping of these objectives to the hypotheses tested in this thesis, see Section B.2.

## 1.5 Primary Contributions

This thesis sits at the intersection of three disciplines: *computer science and high-performance computing* (engineering the infrastructure to process ≈1.7 TiB of multi-site ICU EEG at scale), *data science and machine learning* (domain-adapted feature engineering, modelling, and statistical validation), and *clinical neurophysiology* (ensuring every algorithmic decision reflects what comatose EEG actually represents biologically). The contributions below span all three.

This thesis explicitly separates its scientific findings from the engineering infrastructure required to prove them. The core conceptual and methodological contributions are:

1. **The CN-BAG Methodology:** A novel framework that bypasses the Rejuvenation Artefact by establishing a purely pathological, decision-independent regression baseline.

2. **Domain-Adapted Preprocessing Architecture:** Replacing distribution-relative statistical heuristics with absolute, hardware-physics thresholds, enabling the preservation of critical comatose pathology.

3. **Sex Classification Control:** Establishing binary sex classification as an information-theoretic positive control, providing future researchers with a strict baseline to test whether individual biometric variance survives any proposed data-cleaning pipeline.

4. **State-Trait Inseparability:** Framing the empirical collapse of the predictive models not as algorithmic failure, but as a fundamental neurobiological discovery: acute hypoxic-ischaemic coma completely eclipses underlying biometric identity in the frequency domain.

5. **The WLST Circularity Argument:** A formal Bayesian decomposition demonstrating how withdrawal of life-sustaining therapy structurally inflates the apparent sensitivity of outcome-based prognostic models, providing the methodological justification for abandoning CPC labels as training targets (Appendix D).

## 1.6 Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 details the neurophysiological context of post-cardiac-arrest coma and positions this work against existing brain-age literature. Chapter 3 defines the domain-adapted preprocessing pipeline, the CN-BAG formalisation, and the two-phase experimental methodology. Chapter 4 describes how that design is operationalised against the I-CARE corpus. Chapter 5 presents the empirical evaluation, including the sex classification positive control, the preprocessing impact assessment, and the CN-BAG predictive collapse. Chapter 6 interprets these findings, examining the State-Trait Inseparability Hypothesis and the WLST confound. Finally,

Chapter 7 summarises the primary discoveries and their methodological implications for future neuroprognostication.

# Chapter 2

# Background and Related Work

Building on the motivation and challenges established in Chapter 1, this chapter provides the clinical, neurophysiological, and computational foundations required to evaluate brain age in the comatose ICU environment. It first outlines the intensive care environment and neuroprognostic standards for post-cardiac-arrest patients. It then defines the spectral limits of the coma state, identifies methodological gaps in existing automated preprocessing, and evaluates the brain age paradigm within the context of acute hypoxic-ischaemic injury. Finally, it situates this thesis within the broader literature of machine learning for neurology and signal processing.

## 2.1 Clinical Context and Neuroprognostication

To engineer a valid computational pipeline for ICU EEG, the data cannot be treated merely as an abstract time series; it must be understood within its clinical context. Out-of-hospital and in-hospital cardiac arrests precipitate immediate global cerebral ischaemia. Following successful cardiopulmonary resuscitation and the return of spontaneous circulation (ROSC), patients frequently enter a comatose state driven by severe hypoxic-ischaemic encephalopathy (HIE) [10].

### 2.1.1 Therapeutic Confounds and Clinical Noise

In the critical 72 hours following ROSC, the patient's physiology is actively manipulated by neuroprotective protocols and subjected to the clinical noise of the ICU. These interventions and ambient sources profoundly alter or obscure the biological signals:

- **Targeted temperature management (TTM):** Induced hypothermia ($33^{\circ}\mathrm{C}$–$36^{\circ}\mathrm{C}$)

slows cerebral metabolic rates, mathematically downshifting EEG frequencies and steepening the aperiodic spectral slope [8].

- **Pharmacological Suppression:** Sedatives (e.g., propofol, midazolam) induce highly synchronised anterior alpha and delta oscillations, frequently masking the underlying cortical trait [11].

- **ICU Artefact Profile:** Unlike ambulatory EEG, comatose recordings are heavily contaminated by mechanical ventilation oscillations, infusion pump transients, and involuntary muscle spasms (myoclonus) [12].

### 2.1.2 Terminology Standardisation

In clinical practice, visual EEG interpretation is heavily standardised via frameworks such as the American Clinical Neurophysiology Society (ACNS) Critical Care EEG Terminology. Patterns are categorised into benign (e.g., continuous background, reactivity) and highly malignant (e.g., isoelectric silence, burst-suppression) trajectories.

While clinical guidelines rely on a multimodal prognostic approach, continuous EEG remains the most sensitive real-time monitor of cortical integrity [5]. Recent literature has focused on automating this interpretation using machine learning, achieving high Area Under the Curve (AUC) scores when predicting binary neurological outcomes [13–15].

### 2.1.3 The Label Noise Gap and the Self-Fulfilling Prophecy

While existing predictive models report high statistical accuracy, they suffer from a severe epistemological gap. As detailed in Section 1.1, a significant proportion of mortality in cardiac arrest cohorts follows the withdrawal of life-sustaining therapy (WLST) based on perceived poor prognosis.

This creates a *Label Noise Gap*: if a model is trained directly on survival outcomes, it risks learning to replicate clinician heuristics (and their associated biases) rather than true biological recovery limits. By targeting chronological age—an immutable biological ground truth that is independent of the clinical decision-making process—this thesis bypasses the feedback dynamic inherent in outcome-based modelling.

Empirically, reported false-positive rates for EEG-based mortality prediction after cardiac arrest vary substantially across cohorts and recording sites; Rossetti et al. reported 7 % for unreactive EEG background whilst explicitly acknowledging that a “self-fulfilling

prophecy is probably inherent" in cohorts where clinical parameters inform WLST decisions [16]. Regardless of the specific figure, WLST-driven deaths in the denominator render all such published rates structural lower bounds on the true biological error rate.

## 2.2 Neurophysiology of the Comatose State

Acute HIE induces severe metabolic failure, forcing the brain into critically low-energy states where high-frequency, low-amplitude variance is effectively deleted [17].

### 2.2.1 Neurophysiological Pacemakers: Thalamo-Cortical Loops

The scalp-visible EEG signal reflects the spatially summed postsynaptic potentials of cortical pyramidal neurons, but its rhythmic structure is governed by the *thalamo-cortical loop* [18]. The thalamus acts as the primary pacemaker and gating relay: during relaxed wakefulness, it generates the desynchronised, high-frequency activity of the alert brain and the posterior alpha rhythm (8–13 Hz).

In the context of severe hypoxic injury or pharmacological suppression, these loops fall into pathological synchrony. This transition results in the global dominance of slow oscillations (delta waves, 0.5–4 Hz) and, in extreme cases, the alternating cycle of high-voltage activity and isoelectric silence known as *burst-suppression* [19]. The dominant frequency band of the comatose EEG is therefore not merely a signal feature, but a direct marker of the thalamo-cortical functional state.

### 2.2.2 Feature Typology for EEG Decoding

To translate these pathological states into a predictive matrix, literature traditionally relies on three categories of engineered features:

- *Temporal Features:* Capture waveform morphology and non-stationary transients (e.g., Hjorth parameters, burst-suppression ratios).
- *Spectral Features:* Quantify the distribution of power across the canonical bands defined in Table 2.1, as well as the $1/f$ aperiodic slope.
- *Non-linear Complexity:* Measures such as temporal or spatial Sample Entropy or Lempel-Ziv Complexity, which identify the degree of information density in the cortical signal [20].

- *Source Connectivity:* Graph-theoretical measures such as Weighted Symbolic Mutual Information (wSMI), characterising the functional integration of brain networks.

Combining these features allows for a comprehensive *biometric fingerprint* of the comatose state.

### 2.2.3 Spectral Typology and Frequency Bands

To extract computational features, the EEG signal is decomposed into orthogonal frequency domains. While comatose physiology often blurs these boundaries, the canonical spectral bands defined in Table 2.1 remain the standard for clinical literature.

**Table 2.1:** Canonical EEG Frequency Bands and Comatose Topography

| Band | Range (Hz) | Pathological State Characteristics |
|---|---|---|
| Delta ($\delta$) | 0.5 – 4.0 | Dominant feature of comatose EEG; globally synchronised in deep sedation and severe HIE. |
| Theta ($\theta$) | 4.0 – 8.0 | Often coupled with delta during acute injury; represents diffuse cortical slowing. |
| Alpha ($\alpha$) | 8.0 – 13.0 | Posterior dominant rhythm (PDR) in healthy youth; in coma, often manifests as **Alpha Coma**—a malignant, unreactive, anteriorly distributed pattern [21]. |
| Beta ($\beta$) | 13.0 – 30.0 | Heavily modulated (enhanced) by propofol sedation; serves as a primary pharmacological confounder. |
| Gamma ($\gamma$) | 30.0 – 45.0 | Highly susceptible to electromyographic (EMG) artefact; rarely yields stable comatose pathology. |

### 2.2.4 Signal Acquisition and Montages

The international 10–20 system defines 19 standard scalp electrode positions, providing a spatial map of cortical activity [22]. In a *referential* montage, each electrode is measured against a single reference (e.g., average reference), whereas in a *bipolar* montage, adjacent pairs are differenced. Bipolar montages excel at cancelling common-mode noise but can reduce sensitivity to widely distributed pathological activity. This thesis primarily utilises the 19-channel average reference configuration, which provides the optimal balance for deep learning spatial filtering.

## 2.3 The Brain Age Paradigm and the Rejuvenation Artefact

The Brain Age paradigm relies on training a regressor to map physiological variance to a patient's chronological age. The residual error between the predicted and chronological age—the *Brain Age Gap (BAG)*—serves as a surrogate biomarker for neurological deviation:

$$\text{BAG}_i = \hat{a}_i - a_i \tag{2.1}$$

where $\hat{a}_i$ is the predicted brain age and $a_i$ is the actual chronological age of subject $i$.

The brain's EEG signatures change systematically across the lifespan: *Peak Alpha Frequency (PAF)* decreases by approximately 0.5–1.0 Hz per decade [23], characterized by a flattening of the $1/f$ power spectral density (a reduction in the spectral exponent) [23, 24], and delta/theta power increases relative to the posterior dominant rhythm. A well-trained model captures these co-varying changes as a compressed age prediction.

**The Malignant Alpha Confound.** A critical risk in comatose brain age estimation is the neurophysiological ambiguity of alpha-band activity. While a healthy-normative model interprets a robust 10 Hz oscillation as a *youthful* Posterior Dominant Rhythm (PDR), comatose patients may exhibit *Alpha Coma*—a malignant, unreactive pattern that signifies severe injury [21]. Furthermore, the typically age-associated alpha and delta signals that these models rely upon are often severely attenuated or fundamentally altered in nature by the acute pathology [25, 26]. If a model lacks a coma-normative context, it may systematically misinterpret this pathological sign as a marker of biological youth, necessitating the Coma-Normative approach introduced in this thesis.

### 2.3.1 Obstacles to Domain Transfer

Building on the high-level challenges introduced in Section 1.3, several neurophysiological obstacles prevent the zero-shot transfer of existing brain age frameworks to the comatose ICU state:

1. *Signal domain mismatch:* Burst-suppression and isoelectric intervals are entirely absent from the healthy-EEG corpora used to train standard models.
2. *The Rejuvenation Artefact:* Pathological slowing superficially mimics the high-amplitude delta power of *young* healthy subjects, causing healthy-trained models

to paradoxically underpredict age in severe injury [6].

3. *Aperiodic Masking:* TTM and sedation shift the $1/f$ spectral exponent in directions that physically *invert* the age-related *flattening* reported by Donoghue et al. [24]: whereas ageing progressively reduces the exponent, sedation and hypothermia steepen it, masking or reversing the chronological-age signal.
4. *Cardiac–Injury Coupling:* Cardiovascular dysregulation creates a *Clever Hans* shortcut where models may predict age based on volume-conducted ECG transients rather than cortical dynamics [9, 27].

### 2.3.2 Regression-to-the-Mean Bias

Beyond domain shifts, age regressors suffer from systematic bias where they overestimate the age of young patients and underestimate the age of older patients. Following the frameworks of Liang et al. [28] and de Lange and Cole [29], this thesis applies explicit mathematical bias correction to decouple the prediction error from the target variable.

The *rejuvenation artefact* represents a fundamental transfer barrier. The neurophysiological mechanism is driven by the fact that severe hypoxic-ischaemic injury triggers widespread, high-amplitude delta-wave activity (0.5–4.0 Hz) across the cortex. In healthy ageing, high-amplitude delta is a hallmark of early childhood development and infant sleep; consequently, models trained on healthy norms mistake the pathological slowing of a dying 70-year-old brain for the healthy neuroplasticity of a 7-year-old brain [6]. This confounder fusion means that valid ICU neuroprognostication cannot rely on zero-shot transfer from healthy cohorts, but must explicitly model the comatose state.

### 2.3.3 BAG in Pathological Populations

To anchor the brain age paradigm within a broader medical context, Table 2.2 summarises how BAG varies across diverse pathological states. This pattern establishes the metric as a sensitive, though non-specific, index of neural health.

**Table 2.2:** Brain Age Gap (BAG) in selected clinical populations. Positive values denote apparent age acceleration (older-appearing brain). Sources include both MRI- and EEG-derived metrics.

| Population | BAG Deviation | Outcome Correlates |
|---|---|---|
| Alzheimer's | +2.15–2.47 y | MCI→AD conversion; MMSE decline (MRI-derived) [30] |
| Traumatic Brain Injury | +1.6 y[a] | GCS; post-traumatic amnesia; verbal memory [31, 32] |
| Cardiovascular Risk | +3.5 y | Hypertension; type II diabetes [33] |
| Schizophrenia | +2.6–5.5 y | Grey-matter reduction; bipolar differentiation [34, 35] |
| Epilepsy | +4.5 y | Seizure burden [36] |
| General Ageing | continuous | Life expectancy decreases by 0.81 years and all-cause mortality hazard ratio increases by 1.12 per SD increase in BAG [37] (+6.1% relative increase in risk of death per extra BAG year [38]) |

[a] +6.0 y in prior studies.

## 2.4 Computational Foundations for EEG Decoding

To extract prognostic biomarkers from the comatose signal, the engineering pipeline must resolve the tension between biological variance and technical noise.

### 2.4.1 Signal Properties: Stationarity and Aggregation

Clinical ICU EEG recordings vary drastically in duration and exhibit profound temporal non-stationarity due to fluctuating sedation levels and injury transitions. To resolve this, this thesis relies on the assumption of *piecewise stationarity* [39]. By segmenting the signal into non-overlapping, fixed-duration epochs (e.g., 60 seconds), the signal can be treated as quasi-stationary.

Furthermore, to safely process variable-length recordings, a *hierarchical approach* is required. Local features extracted from independent epochs are aggregated to the patient level (e.g., via medians) before predictive modelling. This prevents *pseudo-replication*, where overlapping windows exhibit strong temporal autocorrelation, and ensures that patients with longer monitoring durations do not statistically overpower shorter recordings during model training.

### 2.4.2 The Preprocessing Paradox

Standard automated toolkits like PREP [40] and Autoreject [41] rely on distribution-relative statistical heuristics. In coma, these frameworks encounter the *preprocessing paradox*: a core tension between the technical requirement to remove noise and the scientific risk of inadvertently deleting high-amplitude pathology. Three specific *Failure Modes* are identified for normative pipelines:

- *Correlation Collapse:* Algorithms like RANSAC assume bad channels exhibit low correlation with their neighbours. However, the hypoxic brain often exhibits global hypersynchrony (e.g., widespread delta waves), causing normative heuristics to falsely flag valid cortical channels as technical noise.
- *Amplitude Threshold Failure:* Standard detectors are calibrated for healthy microvolt ranges ($20$–$50\,\mu\mathrm{V}$). Comatose signals frequently drop into suppression phases below $10\,\mu\mathrm{V}$, causing standard pipelines to indiscriminately delete valid isoelectric periods as *signal loss*.
- *Pathological erasure:* Standard outlier thresholds frequently treat burst-suppression and periodic discharges as non-stationary technical noise, stripping the EEG of its most crucial prognostic features.

### 2.4.3 The Cardiac Fundamental Overlap

A primary driver of the preprocessing paradox is the *Cardiac fundamental frequency* ($\approx 1.2$ *Hz*). This volume-conducted signal overlaps directly with the neural delta band ($0.5$–$4.0$ Hz). Because this overlap is mathematical, aggressive removal of ECG artefacts via spatial filtering (ICA) risks the inadvertent removal of the genuine neural slow waves required for neuroprognostication.

#### Limitations of ASR and ICA

Algorithms like Artefact Subspace Reconstruction (ASR) require an artefact-free baseline segment for calibration, which does not exist in acute coma. Furthermore, because ASR relies on variance-based reconstruction, it frequently misclassifies pathological burst-suppression as a high-amplitude technical artefact, aggressively scrubbing the exact neurophysiological patterns required for neuroprognostication. Similarly, Independent Component Analysis (ICA) coupled with ICLabel [42] often fails to resolve the extensive ECG volume conduction that dominates post-cardiac-arrest recordings.

#### The ECG Confounder and the Minimalist Shift

A primary biomarker for chronological brain ageing is the *flattening* of the $1/f$ aperiodic spectral slope (a reduction in the aperiodic exponent) [23, 24]. However, recent findings demonstrate that age-related changes in $1/f$ dynamics are heavily confounded by cardiac volume conduction in the ICU [9]. This motivates the *quantify-not-remove* strategy of this thesis, and justifies a *minimalist preprocessing shift* for deep learning. As Convolutional Neural Networks (CNNs) can natively learn to bypass artefacts [43], aggressive cleaning consistently reduces decoding accuracy [44].

### 2.4.4 Prediction Reliability and Out-of-Distribution Flagging

In a high-stakes clinical environment, a prediction without a confidence interval is of limited utility. The *fail-safe* requirement mandates that when a recording segment is dominated by artefacts or suppression levels beyond the training range, the pipeline must propagate a low-reliability flag. This ensures that the clinical decision-making process is not misled by overconfident predictions on out-of-distribution data.

### 2.4.5 Deep Learning Constraints for EEG Sequence Modelling

Applying Deep Learning (DL) to clinical electroencephalogram (EEG) data presents certain architectural challenges that distinguish it from standard computer vision or natural language tasks. Primarily, the EEG is a highly non-stationary signal; its statistical properties, such as spectral power and phase coupling, fluctuate dynamically over time in response to both physiological states and pathological injuries [45, 46]. Consequently, networks must be capable of extracting complex, joint dependencies across the spatial, temporal, and frequency domains simultaneously [46].

Furthermore, clinical EEG is typically recorded at high sampling rates (e.g., hundreds to thousands of Hertz), generating exceptionally long data sequences. When applied to these raw sequences, standard Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks face severe computational bottlenecks and are highly susceptible to vanishing gradients [47]. Because processing raw high-frequency EEG with RNNs is computationally prohibitive [47], modern architectures must use sequence compression or alternative temporal mechanisms.

To overcome these constraints, contemporary EEG research frequently favours specific architectural adaptations. Convolutional Neural Networks (CNNs) are widely utilised

as front-end feature extractors to downsample the high-frequency time-domain signal into a condensed, multi-scale latent representation [48]. For sequence modelling, Temporal Convolutional Networks (TCNs) use causal, dilated convolutions to establish an extended receptive field without the recursive penalties of RNNs [49]. Alternatively, Transformer networks bypass recurrence entirely, utilising self-attention mechanisms to efficiently map long-range temporal dependencies and dynamically weight the most relevant non-stationary features across the recording [49, 50].

## 2.5 Algorithmic Architectures for EEG Decoding

### 2.5.1 Tree-Based Ensembles and Gradient Boosting

For tabular feature matrices, gradient-boosted trees remain the industry standard for interpretability and predictive performance. Algorithms such as XGBoost [51], Random Forests, and ElasticNet serve as primary benchmarks with complementary inductive biases. XGBoost builds ensembles of weak learners sequentially, where each new tree corrects the residual errors of the previous ensemble [52]. Its native handling of missing values makes it well-suited to ICU EEG features, where quality masking produces structured missingness.

### 2.5.2 Deep Learning and Sequence Modelling

End-to-end deep learning bypasses manual feature engineering by learning optimal spatial and temporal filters directly from the signal tensor.

- *Convolutional Neural Networks (CNNs):* Architectures like EEGNet [53] use 1D convolutions to extract waveform transients and local temporal patterns at scales from milliseconds to seconds.

- *Recurrent Networks (RNN/LSTM):* Long Short-Term Memory (LSTM) networks model temporal dependencies via gated cell states. In ICU monitoring, CNN-BiLSTM-Attention architectures provide a multi-scale solution: CNN layers detect local waveform features, the BiLSTM captures longer-range rhythmic context (e.g., burst-suppression cycles), and attention mechanisms provide soft temporal weighting for non-stationary biological arrays [49].

- *Sequence Modelling Mechanisms:* Temporal Convolutional Networks (TCNs) or self-attention Transformers provide stable, wide receptive fields, though they incur higher computational costs for high-frequency EEG sequences [54].

### 2.5.3 Interpretability and Shortcut Learning

To meet the clinical requirement for transparency, the model's decision-making process must be accessible. For feature-based models, SHapley Additive exPlanations (SHAP) are used to attribute predictions to individual input features via game-theoretic marginal contributions [55]. For deep learning architectures, attention weights offer a complementary window into which temporal segments were most informative.

These interpretability safeguards are essential to detect *shortcut learning*, where models exploit peripheral physiological signals—specifically volume-conducted ECG—as a shortcut for age prediction. This phenomenon is severe in ICU EEG, where the $1/f$ aperiodic slope is substantially driven by cardiac conduction rather than cortical dynamics [9, 27].

However, applying these tools to severely pathological populations introduces a hard methodological constraint. Feature attribution methods mathematically assume that the model has successfully learned a valid, variance-explaining mapping of the underlying biology. If a model fails to overcome the domain transfer barriers outlined in this chapter—collapsing instead into a constant predictor with a coefficient of determination near zero ($R^2 \approx 0$)—any extracted feature attributions become mathematically meaningless. Consequently, the two-stage strategy of this thesis ensures that an interpretable, feature-based baseline establishes a valid predictive floor before attempting to interpret higher-order deep learning representations.

## 2.6 The Landscape of Large-Scale EEG Corpora

The development of generalisable machine learning models requires large, heterogeneous datasets. Several benchmark corpora have emerged, providing the normative distributions against which clinical deviations are measured. Table 2.3 provides a comprehensive overview of these principal datasets, sorted by their relevance to the primary research domain of this thesis, post-cardiac-arrest coma, detailing participant numbers, sex distributions, and age ranges.

Datasets vary largely in quality (sample rate, number and montage of channels,

**Table 2.3:** Overview of large-scale EEG and MEG datasets sorted by relevance to post-cardiac-arrest comatose patient research. Missing demographic data indicates aspects not fully detailed in the primary sources.

| Dataset | Subjects ($N$) | Age (Years) | Sex (F/M) | Modality & Clinical Context |
|---|---|---|---|---|
| HEEDB | 108,341 | 0–90+ | 49% / 51% | Routine, EMU, and ICU EEG; highly diverse clinical archive [56] |
| SHHS | 5,793 | 40–90 | 52% / 48% | Overnight PSG/EEG; healthy ageing and sleep disorders [57] |
| MGH Sleep Laboratory | 8,673 | 18–90+ | — | Overnight PSG/EEG; healthy adult sleep registry; brain-age subset: $N = 2{,}532$ [33] |
| TUH (TUAB Subset) | 2,329 | 0–95 | 54% / 46% | Routine Scalp EEG; mixed normal and abnormal pathologies [58] |
| TDBRAIN | 1,274 | 5–88 | 49% / 51% | Resting-state EEG; high psychiatric and cognitive burden [59] |
| LEMON | 227 | 20–35; 59–77 | 36% / 64% | Resting-state EEG & multimodal MRI; healthy bimodal age design [60] |
| I-CARE | 1,020[b] | 16–90+ | 33% / 67%[c] | Continuous ICU EEG; post-anoxic coma and severe HIE [61, 62] |

[b] 607 published. [c] For the published subset.

accompanying annotation or metadata, additional non-EEG channels) and recording length (spanning from 1 minute to continuous EEG of hundreds of hours). For instance, the I-CARE 607 published patients make $\approx 1.7$ TiB in size, while the TUH full dataset corpus with data from more than $15,000$ patients has a size of $\approx 0.6$ TiB. Most larger corpora are available for scientific projects after registration.

**Clinical and Hospital Archives.** The Temple University Hospital (TUH) EEG Corpus [58] is the most established clinical resource (with over $70,000$ recordings from $N > 15,000$ subjects). It offers various corpora for abnormal and pathological (e.g. epilepsy) data and is partially annotated. The *Harvard Electroencephalography Database (HEEDB)* aggregated over $280,000$ routine and ICU recordings from $N = 108,341$ patients at the Boston hospitals [56, 63]. Access to both datasets is restricted.

**Research and Healthy Cohorts.** Datasets such as TDBRAIN ($N = 1,274$) provide resting-state EEG primarily for patients with psychiatric conditions (e.g., major depressive disorder), containing a very limited subset of healthy controls ($N = 47$) [27, 59, 64]. Pristine research-grade cohorts like *Cam-CAN* [65] offer insights into ageing trajectories across the lifespan but utilise magnetoencephalography (MEG) rather than EEG, and lack the severe clinical noise required to train models robust to the ICU [65]. The *Leipzig Mind–Brain–Body (LEMON)* dataset [60] provides a compact healthy bimodal cohort ($N = 227$; young: 20–35 years, $N = 153$; elderly: 59–77 years, $N = 74$) with concurrent 62-channel resting-state EEG and multimodal MRI, but its small size and deliberate two-group design limit its suitability as a training corpus for continuous age regression.

**Sleep and Polysomnography Cohorts.** The MGH Sleep Laboratory (full registry: $N \approx 8{,}673$) and the *Sleep Heart Health Study (SHHS, $N = 5{,}793$)* [57] provide thousands of overnight polysomnography recordings spanning the adult lifespan. Brain-age models trained on these sleep cohorts have achieved MAE of 7.6 years [33]; deep sequence learning further reduces this to MAE 4.19 years via Swin Transformer architectures [66]. However, these cohorts rely on sleep micro-architecture (e.g., spindles, K-complexes) which is fundamentally absent in post-anoxic coma, limiting direct transfer.

**The I-CARE Database and the Domain Gap.** The I-CARE database ($N = 1,020$) remains unique in its focus on post-cardiac-arrest ICU patients [61]. Unlike routine clinical archives, the I-CARE arrays are dominated by pathological hypersynchrony

(burst-suppression), targeted temperature management distortions, and extensive volume-conducted ICU noise.

## 2.7 Literature Synthesis and Research Gap

The existing literature on EEG brain age estimation has evolved into four primary thematic clusters, ranging from foundational biomarkers to the identification of critical gaps in clinical transfer.

**Foundations and Biological Markers.** Early work established the primary spectral correlates of biological ageing, specifically the reduction in alpha peak frequency and the *flattening* of the $1/f$ aperiodic slope [23, 24]. These markers provide the physiological justification for viewing the EEG as a proxy for neural health.

**Benchmarks in Healthy and Sleep Cohorts.** Large-scale studies on awake resting-state and sleep EEG have set the current performance ceiling for the field, achieving Mean Absolute Errors (MAE) of 4.2 to 8.0 years [33, 67]. These studies serve as the *ideal state* benchmarks against which clinical populations are measured.

**The Debates: State-Trait and Shortcut Learning.** Recent literature has highlighted key risks in normative modelling. The *state-trait* debate evaluates whether acute pathological states (like anaesthesia or coma) permanently obscure the underlying biological trait [6, 68]. Concurrently, the *Clever Hans* or shortcut learning debate warns that models may be learning peripheral physiological signals (specifically ECG) as a shortcut for age prediction [9, 27].

**The Coma Gap.** Despite these advancements, a significant methodological void remains: no published study has successfully applied the brain age paradigm to a post-cardiac-arrest comatose cohort. Existing pipelines are either too destructive (deleting pathological variance) or too dependent on contaminated outcome labels. This thesis addresses this gap through the CN-BAG framework.

### 2.7.1 Comparative Analysis

In healthy resting-state EEG, models achieve a Mean Absolute Error (MAE) of 6–8 years [67]; sleep-recorded cohorts yield 4.19–7.6 years depending on architecture [33, 66].

Table 2.4 compares primary EEG brain age studies against the proposed framework. The consistent pattern across the literature is that performance degrades as the distance from healthy resting-state increases. Notably, reported $R^2$ values in standard pipelines substantially overstate true cortical prediction performance: Bomatter et al. showed that ICA removal of peripheral (cardiac, muscular, ocular) signals reduces $R^2$ by $0.046$ on average ($p = 0.008$), directly demonstrating that a meaningful fraction of apparent brain-age accuracy is a peripheral biomarker result[1] [27]. No published study has successfully navigated the *state-trait eclipse* of the acute comatose state, marking a significant methodological void in the field.

This chapter has established the foundations required to operationalise the CN-BAG biomarker. Chapter 3 translates these into the scientific design: the preprocessing approach, feature extraction, and modelling strategy.

[1] This finding directly motivates the *EEG ECG Contamination Index* (EECI) proposed in Section 3.3.8: rather than blindly removing cardiac signals and losing signal, the EECI quantifies contamination severity so it can be used as a covariate.

**Table 2.4:** EEG brain age studies related to this thesis. Reports the best lowest MAE and highest $R^2$ achieved. *Norm:* H = healthy population (external), P = population-normative (within-distribution), H+P = combined. *State:* Aw = awake, Sl = sleep, An = anaesthesia, Mi = mixed, Co = coma.

| Study | Dataset(s) | $N$ | Norm | MAE | $R^2$ | Key limitation | State |
|---|---|---|---|---|---|---|---|
| Zhang et al. [66] | SHHS, MROS, WSC, NCHSDB, … | 13,616 | H+P | 4.19 y | 0.94[d] | Cohort variation; limited clinical metadata | Sl |
| Engemann et al. [67] | TUAB, LEMON, CHBP, Cam-CAN | 2,540 | H | 7.30 y[f] | 0.74[f] | Demographic bias; montage dependency | Aw |
| Gemein et al. [6] | TUAB | 2,329 | H+P | 6.6 y | 0.73 | State–trait; pathological training data | Mi |
| Sun et al. [33] | SHHS, MGH | 4,506 | H | 7.6 y | 0.67[e] | PSG montage | Sl |
| Sabbagh et al. [68] | Unspecified (OR) | 323 | H | 8.2 y | 0.65 | *OR only*; elective; propofol | An |
| Bomatter et al. [27] | TUAB, TDBRAIN | 2,494 | H+P | — | 0.4–0.8[g] | Cross-sectional; artefact dependent | Aw |
| **This thesis** | **I-CARE** | 604 | **P** | — | — | **ICU** | **Co** |

[d] Derived from $r = 0.97$. [e] Derived from $r = 0.82$ (Pearson correlation). [f] Cam-CAN MEG subset (highest $R^2$ across all conditions); best EEG-only MAE was 6.48 y on CHBP with $R^2 = 0.18$ only.

[g] Pipeline- and dataset-dependent (TUAB ≈0.4–0.6, TDBRAIN ≈0.8 with autoreject). ICA removal of peripheral artefacts reduces $R^2$ by $-0.046$ on average ($p = 0.008$), directly demonstrating cardiac and muscular signal exploitation. See Section 3.3.8.

# Chapter 3

# Methodology

This chapter translates the theoretical constraints of comatose neurophysiology established in Chapter 2—in particular the State-Trait Inseparability problem, the neurophysiological dynamics of post-cardiac-arrest coma [69, 70], and the domain-shift limitations of healthy-normative brain age models [6]—into a concrete, well-controlled study design. This chapter follows the data path: beginning with design philosophy and experimental architecture, proceeding through preprocessing, feature engineering, and modelling, and closing with dataset description, scientific integrity, and project governance.

1. **Design Philosophy (Section 3.1):** The definition of CN-BAG and the overarching goal-driven engineering principles.

2. **Experimental Design (Section 3.2):** The complete experimental structure, standardised data splits and the sex classification control design.

3. **Preprocessing Architecture (Section 3.3):** The domain-adapted signal preprocessing and quality masking.

4. **Feature Engineering (Section 3.4):** The handcrafted feature typology and hierarchical aggregation.

5. **Experimental Tracks (Sections 3.5 and 3.6):** The Feature-based Machine Learning (Phase 1) and Time-Frequency Signal Deep Learning (Phase 2) experimental tracks.

6. **Dataset and Infrastructure (Section 3.7):** The I-CARE dataset and HPC orchestration.

7. **Scientific Integrity and QA (Section 3.8):** Explorative Data and Validity Analyses, Statistical Planning, and Evaluation.

8. **Project Governance and Management (Section 3.9):** The project methodology, SSOT, and workflows.

To facilitate direct comparison between the scientific intent and reasoning and the empirical execution and results, the thesis maintains a vertically aligned structure across chapters, as displayed in Table 3.1. While the focus clearly shifts towards the experimental implementation (Chapter 4) and the subsequent reporting of results (Chapter 5), structural changes are kept at a minimal level. Additional technical detail is consolidated in the appendices and cross-referenced from the core chapters.

**Table 3.1:** Structural alignment of the "Scientific Why" (Methodology), "Experimental How" (Experiments), and "Empirical Result" (Results) across the core chapters, supplemented by the technical appendices.

| Analytical Stage | Methodology | Experiments | Results | Appendix |
|---|---|---|---|---|
| Design Philosophy | Section 3.1 | — | Section 5.1 | Appendix D |
| Experimental Design Architecture | Section 3.2 | Section 4.2 | — | Sections B.4 and B.5 |
| Preprocessing | Section 3.3 | Section 4.3 | Section 5.2 | Sections A.1 and A.1.1 |
| Feature Engineering | Section 3.4 | Section 4.4 | Section 5.3 | Sections A.4 and A.5 |
| Phase 1 (Feature-Based ML) | Section 3.5 | Section 4.5 | Section 5.4 | Sections B.10 and B.11 |
| Phase 2 (Deep Learning) | Section 3.6 | Section 4.6 | Section 5.5 | Section B.12 |
| Dataset and Infrastructure | Section 3.7 | Section 4.7 | — | Sections C.2 and C.2.1 |
| Scientific QA | Section 3.8 | Section 4.8 | Section 5.6 | Appendix E |
| Project Governance | Section 3.9 | Section 4.9 | — | Sections C.1 and C.1.2 |

## 3.1 Design Philosophy

### 3.1.1 Definition of the Coma-Normative Brain Age Gap (CN-BAG)

Defining the study's central biometric is the first step in specifying the analytical objective. Extending the generic Brain Age Gap definition (Equation (2.1)), the *Coma-Normative*

*Brain Age Gap (CN-BAG)* is operationalised using the coma population as its own normative reference:

$$\text{CN-BAG}_i = \hat{a}_{\text{coma}}(\mathbf{x}_i) - a_i \tag{3.1}$$

where $\hat{a}_{\text{coma}}$ is the predicted age based on the comatose EEG representation $\mathbf{x}_i$, and $a_i$ is the chronological age (chosen for its independence from clinical treatment decisions; see Section 2.1.3). By using the coma population as its own normative reference, the framework absorbs the shared pathological variance (e.g., sedation effects) to isolate individual-level traits.

The primary target is deliberately conservative: the later 12-year MAE threshold is a feasibility objective, not a deployment claim, chosen to sit above trivial performance while remaining realistic under the severe healthy-to-coma domain shift. Concretely, a constant predictor that always outputs the cohort mean yields an MAE of approximately 12.6 y for the training cohort ($\sigma_{\text{age}} \approx 15.6$ y, $N = 386$); any model that fails to beat this threshold provides no information beyond the population average.

This approach deliberately avoids the *Rejuvenation Artefact* encountered when applying healthy-trained models to pathological data. For instance, the CN-BAG framework resolves the *alpha coma paradox*: in a healthy-normative model, a robust 10 Hz alpha oscillation is interpreted as a youthful trait; in a coma-normative model, the same oscillation is identified as a marker of severe pathological deviation.

**State-Trait Inseparability: The Negative Motivation.** The central challenge of this study is the **State-Trait Inseparability** problem. The observed EEG signal $\mathbf{y}_i(t)$ is modelled as a compound of the following components:

$$\mathbf{y}_i(t) = j(f(a_i), g(s_{i,t}), h(b_{i,t}), \nu(n_{i,t}), \varepsilon_{i,t}) \tag{3.2}$$

where:

$j(\cdot)$**:** The complex, non-linear mixing function combining individual components into the observed scalp signal.

$f(a_i)$**:** The target *Age Trait* (chronological age $a_i$).

$g(s_{i,t})$**:** The dominant *State Effect* (hypoxic injury, sedation, metabolic encephalopathy).

$h(b_{i,t})$**:** Individual *Physiological Noise and Biometrics* (physiological artefacts, skull thickness, electrode impedance, baseline physiology).

$\nu(n_{i,t})$**:** Environmental *Technical Noise* (ICU environment, amplifier saturation).

$\varepsilon_{i,t}$**:** Random noise *Residual Variance*.

The individual components and their respective effects and interaction with j are unknown, but age trait $a_i$ must be assumed to resemble the correct chronological age with limited variance. The central risk governing this design is that the magnitude of the state effect $\|g(s_{i,t})\|$ eclipses the age-related variance $\|f(a_i)\|$ in their interaction in coma ICU. In standard normative paradigms, this leads to the *Rejuvenation Artefact*, where comatose patients are systematically predicted as younger than their chronological age because widespread high-amplitude delta activity mimics the spectral signature of young healthy brains rather than the age-related electrophysiological signature of the patient's true cohort. The CN-BAG design mitigates this by absorbing the effects of $g(s_{i,t})$ into the population-level baseline, forcing the model to learn the effects of $f(a_i)$ that survive the pathological eclipse through the aggregating function $j$.

### 3.1.2 Central Design Commitments

The methodology is organised around three central commitments. First, the target is defined relative to the coma population itself, because both pathological state and pharmacological context can shift the electrophysiological baseline enough to make healthy- or alternate-state transfer systematically misleading [6, 68, 71]. Second, preprocessing is conservative and coma-adapted: only physically impossible signal is removed, while physiologically plausible pathological structure is retained and annotated, since burst-suppression and related abnormal patterns can themselves carry prognostic or age-related information [19, 27, 72]. Third, evaluation proceeds in two stages under a pre-specified analysis plan: an interpretable baseline first tests whether trait information survives at all, and a sequence model then tests whether additional temporal structure can be recovered when time-varying coma dynamics are modelled explicitly [67, 70, 73]. Thresholds and profile settings are fixed on the development data before full-cohort execution; grouped validation and empirical null checks are likewise pre-specified to limit leakage and optimistic inference [74, 75]. Chapter 4 describes that operationalisation.

## 3.2 Experimental Design

The study is structured into a preparatory phase outlined above and in two experimental phases:

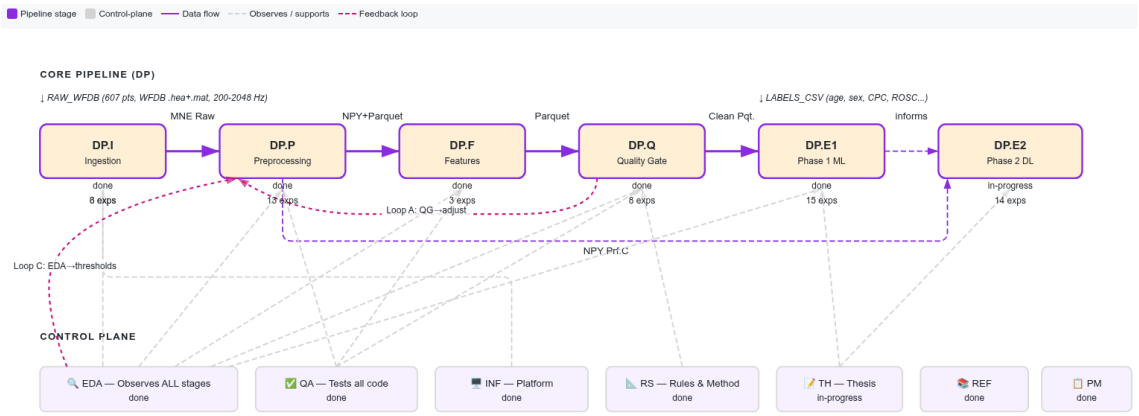



**Figure 3.1:** Study pipeline and control plane overview. Raw EEG flows left to right through preprocessing, feature extraction, and dual-track modelling (Phase 1 / Phase 2) into evaluation. The *Control Plane* (top) represents cross-cutting functions — quality assurance, exploratory analysis, infrastructure, and project governance — that interact with the data flow at every stage; the quality gate loop indicates that preprocessing and features are reviewed iteratively before experiments are finalised. Central data formats at each stage transition are shown.

- *Phase 0 (Preprocessing and Feature Extraction):* Establishes the basis for the non-destructive signal preprocessing.

- *Phase 1 (Feature-Based Classical ML):* Targets *physiological validation* (the *Why*). By using hand-crafted features, this phase asks whether a recoverable age signal exists at all and whether it is anchored in cortical oscillations rather than in artefact.

- *Phase 2 (End-to-End Deep Learning):* Targets *predictive performance* (the *Max*). This phase asks whether latent temporal dynamics, invisible to handcrafted extractors, can improve on the Phase 1 floor.

Figure 4.1 (in Chapter 4) shows the modular data flow and central integration with the *control plane*, while Figure 3.2 depicts the full data processing architecture. Crucially, these phases are architectural decisions rather than contingency branches: Phase 1 provides the statistical anchor, while Phase 2 provides the representation-learning ceiling. This *Dual-Track* approach ensures that even if Phase 2 achieves superior power, its biological validity is anchored by the interpretable Phase 1 findings. The final test set is only evaluated on the Phase 2 best performing model.

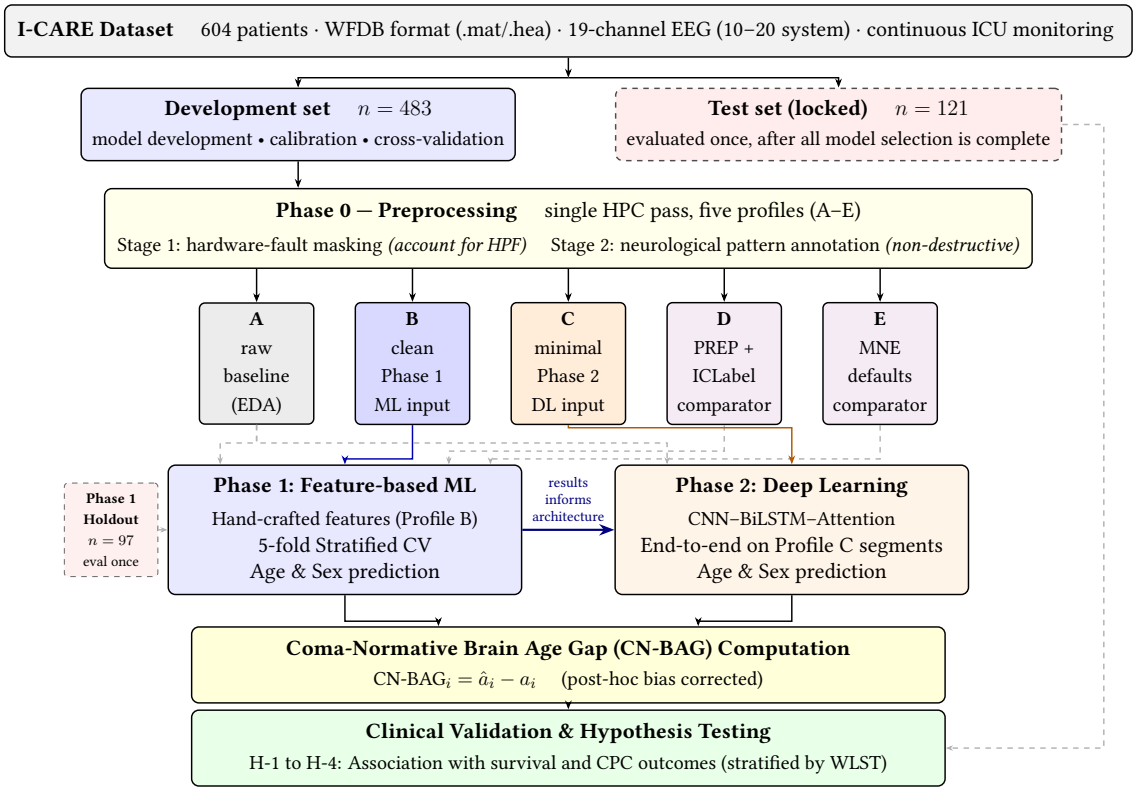



**Figure 3.2:** The full processing pipeline is engineered to uphold scientific integrity and reproducibility, employing strict quality-first, domain-knowledge principles. Gated experimental design overview. The design enforces subject-level isolation from raw data through to outcome association. **Phase 0** provides the shared preprocessing infrastructure including feature extraction. **Phase 1** targets baseline identification with physiological validation via handcrafted features. **Phase 2** tests the representation ceiling via end-to-end deep learning. The final evaluation (H-1 to H-11) occurs only after the test set is unsealed.

### 3.2.1 Integrating the Data Plane: From Signal to Gap

The scientific design is operationalised through a seamless integration of preprocessing, feature engineering, and normative modelling. The architecture enforces a strict non-destructive sequence: raw waveforms are recovered via domain-adapted filters, masked for hardware integrity, and reduced to stable patient-level traits before being ingested by the CN-BAG models. The complete experiment dependency graph, tracing every EXP-ID through its quality gates and inter-experiment links, is provided in Section B.4.

This integration ensures that every sample-level quality decision (the "How") is auditable and linked directly to the biometric recovery objective (the "Why"). To visualise the data flow across these stages, the design distinguishes between the *Control Plane* (how subjects are partitioned) and the *Data Plane* (how waveforms are routed through specific algorithmic derivations).

### 3.2.2 Standardised Data Partitions and Diagnostic Design

Data leakage was prevented through three nested isolation constraints applied to the 1.7 TiB corpus before any modelling began:

1. **Subject-Level Isolation (The Sealed Test Set):** The primary partition splits the $N = 604$ patients into a **Development Set ($n = 483$)** and a strictly isolated **Sealed Test Set ($n = 121$)** as shown in Figure 3.3. All feature selection and hyperparameter tuning are performed exclusively within the development set.

2. **Diagnostic Temporal Holdout:** Within the development set, a nested temporal split is implemented to detect *temporal pseudo-replication*. For each patient, the final $20\%$ of their recording duration is reserved for validation, ensuring that the model learns generalisable biological traits rather than idiosyncratic, time-varying state markers, while also confirming that the model is capable of learning from the comatose EEG representation at all—even if only within the training segment.

3. **Sex Classification Control:** Before evaluating the subtle age signal, the pipeline must successfully recover biological sex—a trait with a known, robust EEG signature. Failure at this *positive control* would establish the information-theoretic floor of trait erasure in the coma state (the empirical results for this test are documented in Chapter 5).

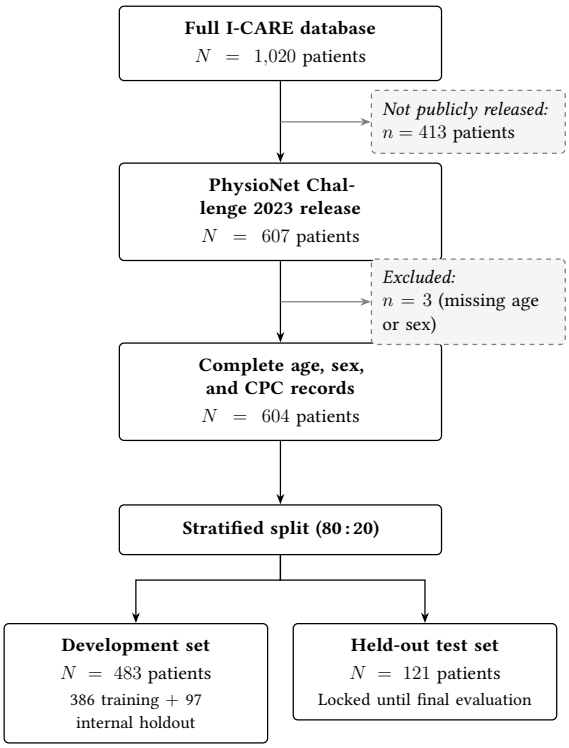



**Figure 3.3:** CONSORT-style cohort flow diagram. The diagram illustrates the systematic reduction from the global 1,020-patient I-CARE repository to the 604-patient high-integrity cohort used for modelling. **Exclusion criteria** are restricted to technical incompleteness (missing age/sex labels) to preserve a representative pathological sample. The resulting $N = 604$ cohort is partitioned into a development set ($n = 483$) and a sealed test set ($n = 121$) to ensure strict evaluation integrity.

## 3.3 Preprocessing Architecture

### 3.3.1 Beyond Standard Toolkits

The preprocessing design is summarised against the most common alternatives in Table 3.2 to make the methodological trade-off explicit.

**Table 3.2:** Summary comparison of this work against common normative preprocessing stacks. The goal is not feature parity, but coma-specific threshold anchoring and burst-suppression preservation.

| Dimension | This work | PREP | MNE / ICLabel |
|---|---|---|---|
| Threshold anchoring | Absolute hardware bounds and quality mask | Distribution-relative heuristics | Default/heuristic settings |
| Burst-suppression handling | Retained and annotated unless physically impossible | Typically treated as artefact risk | Often pruned or under-modelled |
| Coma adaptation | Explicit coma-normative profiles A–E | Healthy-EEG assumption | General-purpose EEG defaults |
| Cardiac contamination | Quantified, then selectively removed or retained by profile | Not primary design target | Requires extra custom logic |

Two additional, partially contradicting biophysical constraints dictate the architectural choices:

- **Non-Stationarity** ICU EEG is not a stable signal; it exhibits rapid state-transitions (e.g., burst-suppression cycles, telemetry dropouts). This mandates a *granular quality assessment* on short intervals.
- **Long-Dependencies:** The relevant neurophysiological traits (e.g., connectivity, fractal complexity) require windows long enough to capture low-frequency delta stability ($< 1$ Hz), motivating the Feature Engineering as outlined in Section 3.4.

### 3.3.2 Phase 0: Shared Infrastructure

Crucial to the study design is the **Phase 0 Shared Modular Infrastructure**. Rather than building two separate pipelines for classical ML and deep learning, the study implements a

unified *Data Plane* that provides a standardised, high-integrity signal recovery foundation for all subsequent modelling tracks.

This shared infrastructure enforces three universal constraints:

1. **Hardware Integrity First:** Both Phase 1 and Phase 2 ingest waveforms that have been passed through the same `uint8` Quality Mask (Section 3.3.5), ensuring that no model—regardless of its complexity—is exposed to amplifier saturation or electrode-fault artefacts.

2. **Standardised Resampling:** All signals are downsampled to a uniform 256 Hz via a Kaiser-windowed FIR filter to prevent aliasing, providing a consistent temporal resolution across multi-centre hardware configurations.

3. **Profile-Based Branching:** The infrastructure supports a fixed set of preprocessing profiles (A–E). While loading and masking are shared, the pipeline branches into phase-specific derivations only after those common steps are complete.

By centralising these operations, the design reduces the risk that performance differences between Phase 1 and Phase 2 simply reflect inconsistent upstream cleaning rather than differences in representation.

### 3.3.3 Multi-Profile Preprocessing Baseline

To empirically evaluate the impact of preprocessing on prognostic modelling and test the clustered hypotheses (e.g., on data quality impact), the experimental design routes the raw data through five discrete preprocessing profiles (A–E). This comparative design allows isolation of the biometric "Why" from the technical "How." The multi-profile comparison also makes the standard-toolkit baseline explicit; see Table 3.2.

- *Profile A (Raw baseline):* Mandatory Stage 1 loading, quality masking, and resampling only. Acts as the information-theoretic reference.

- *Profile B (Bipolar, domain-adapted):* Specifically targeted for Phase 1 classical ML; uses bipolar derivation to reject cardiac noise while preserving delta-band power.

- *Profile C (Minimal deep-learning baseline):* Tailored for Phase 2 deep learning; uses common-average referencing with minimal filtering while intentionally retaining ECG contamination in the waveform so the sequence model sees the full tensor stream.

- *Profile C-noecg (Cardiac-removal ablation):* Matched to Profile C but adds targeted ECG-ICA, allowing the effect of cardiac component removal to be tested explicitly rather than assumed.
- *Profile D (Normative Comparator):* Industry-standard healthy-EEG pipeline (PREP/RANSAC) used as a negative control to quantify failure modes on pathological data.
- *Profile E (MNE Defaults):* Basic minimal preprocessing using standard off-the-shelf software defaults, included to benchmark whether the domain-adapted pipeline introduces systematic biases relative to a generic, unconfigured baseline.

Full step sequences and parameter settings for each profile are given in Section A.1.1; the routing diagram is shown in Figure 4.7 in Chapter 4.

### 3.3.4 Conservative Handling of Pathology and Artefact

One of the central methodological problems in comatose EEG is the absence of reliable ground truth for separating artefact from pathology at scale. Severe brain injury can produce globally synchronised, discontinuous, or suppressed patterns that off-the-shelf toolkits treat as noise, even though those same patterns are clinically informative in post-anoxic coma [69, 72, 76].

For that reason, the pipeline removes only data that is physically impossible or clearly hardware-derived. Other biological variation is retained and annotated, even when it is clinically abnormal or potentially contaminated. This keeps the preprocessing conservative and allows downstream analyses to test whether cardiac, muscular, or burst-suppression-related structure is informative rather than assuming that it should be deleted a priori [27, 77, 78]. For the Phase 2 deep learning track, this design is additionally supported by the established capability of end-to-end convolutional architectures to suppress non-neurogenic structure without explicit removal: artefact rejection has been shown empirically not to improve CNN-based classification performance on clinical EEG, with networks adapting their internal representations to discount structured non-neural signals when trained end-to-end [6, 43].

### 3.3.5 The Masking-First Principle

The sequence of preprocessing operations is strictly ordered to prevent the *artefact-smearing* effect. The *Quality Mask* is generated before any spectral filtering so that

amplifier spikes, flatlines, and related hardware faults are not spread into adjacent clean signal by later filtering steps. The step-by-step ordering and the rationale for each branch are summarised in Section A.1.4.

### 3.3.6 Quality Mask Engineering

The signal recovery is operationalised via a two-tier masking design that separates hardware faults from post-filter physiological or pathological annotations. Stage 1 identifies non-biological failure states in the raw signal before filtering; Stage 2 records residual artefacts and neurological patterns as features after filtering without deleting them. In Phase 1 this design determines which windows remain analysable, whereas in Phase 2 the same information from Stage 1 is available to the model as structured mask input to determine higher-quality segments. The detailed bitfield schema, thresholds, and mask examples are given in Sections A.2 and A.2.1.

### 3.3.7 Spectral Cleaning: Prominence-Gated Notch and Zapline

Spectral integrity is maintained via a two-tier strategy. The primary mechanism is a *prominence-gated IIR notch*: the pipeline measures the spectral peak prominence at the WFDB-header line frequency (50 or 60 Hz), bypasses the notch entirely when the peak is below 3 dB, and otherwise applies a second-order Butterworth notch with an adaptively estimated Q-factor (clamped to $[20, 80]$; see Section A.1.2 for the resolution limitation of the Q estimator). The IIR filter type was selected over FIR by a cohort-wide patient-level vote (EXP-306). For profiles requiring the highest spectral fidelity, the pipeline alternatively employs **Zapline** (DSS-based spectrum-subtraction), which exploits the spatial redundancy of the multi-channel EEG array to identify and subtract line-noise components without creating the spectral discontinuities typical of notch filters.

The filter behaviour is illustrated in Section A.1.2.

### 3.3.8 The EEG-ECG Contamination Index

A critical risk in this domain is that brain age models may inadvertently measure the heart rather than the brain. Cardiac volume conduction embeds the QRS complex and its harmonics throughout the EEG spectrum, overlapping the 1–3 Hz delta band that carries the dominant coma-state signal. Standard artefact removal via Independent Component Analysis (ICA) can separate the cardiac component in healthy EEG, but in severely

suppressed comatose recordings it frequently destroys up to 50% of genuine neural delta power, removing Heartbeat-Evoked Potentials alongside the artefact. Schmidt et al. showed that the 1/f aperiodic slope is substantially driven by cardiac volume conduction and that standard ICA removal *reduces* age-prediction accuracy [9]; Bomatter et al. demonstrated that uncontrolled contamination leads models to exploit peripheral rather than cortical dynamics [27] (see Sections 2.7.1 and 6.2).

To our knowledge, no validated composite index simultaneously quantifies this contamination across amplitude, spectral-coherence, and phase-synchronisation domains for simultaneous EEG–ECG recordings in clinical settings [12]. The literature remains confined to single-domain approaches (ICA-based subtraction, frequency-band regression, adaptive filtering), none of which yield a per-channel, per-window contamination *estimate* suitable for downstream covariate use.

The *EEG-ECG Contamination Index (EECI)* is proposed as a composite metric addressing this gap, following a *quantify-before-removal* philosophy: rather than blindly subtracting suspected artefact, the metric measures contamination severity so that it can be accounted for as a comprehensive statistical covariate. The proposed score draws on multiple domains: amplitude (cross-correlation with the ECG channel), spectral coherence in the heart-rate and QRS transient bands, Phase-Locking Value to the cardiac cycle, and Mutual Information. The precise weighting of these components and resolving their interdependencies require theoretical modelling and empirical calibration, as the relative contributions to cortical signal are unknown.

This objective was deferred as a primary research aim. The sequencing reflects a deliberate priority: it is necessary first to establish whether a brain age signal survives the comatose state before investing in the formal calibration infrastructure the EECI requires. Calibrating the metric's weights and thresholds further requires a clinical neurophysiology collaborator to annotate a representative set of ICU EEG windows—a task outside the present thesis timeline. The prototype was implemented; the design rationale, component specification, and calibration pathway are documented in Section B.2.4. Crucially, the absence of any standardised multi-domain cardiac contamination index in the wider literature means that the EECI, once validated, represents a self-contained methodological contribution independent of the CN-BAG framework itself.

### 3.3.9 Patient-Level Robust Normalisation

Normalisation is used only where the model class requires it. While coarse centring is handled during signal loading, Phase 2 applies patient-level robust scaling (median/IQR) to the tensor input so that cross-centre gain differences do not dominate optimisation.

*Segment-level* normalisation was rejected because it destroys relative amplitude dynamics between windows, and *batch-level* normalisation was rejected because it introduces training instability. Scaling each patient's data by that patient's own median and IQR preserves the internal dynamics of the stay while keeping amplitudes within a stable range for sequence modelling.

Phase 1 employs raw-scale, fully interpretable features and intentionally omits normalisation beyond profile-specific signal montages. In this phase, absolute signal amplitude is treated as a primary feature of interest, as its preservation is critical for capturing pathological markers like voltage attenuation. Robustness is ensured through domain-specific feature design and patient-level aggregation. While Phase 1 makes this trade-off in favour of interpretability, Phase 2 receives signals that have already undergone patient-level robust normalisation (median/IQR per channel, applied during preprocessing) to mitigate cross-centre gain differences and stabilise training-target convergence.

### 3.3.10 Hardware-Grounded Thresholds

The preprocessing approach therefore avoids thresholds borrowed unchanged from healthy-EEG defaults and instead uses hardware-grounded limits that can be justified independently of downstream model performance. The external literature does not present this as a single standard ICU recipe; rather, it repeatedly shows that distribution-relative or spatially heuristic cleaning can be counterproductive in abnormal EEG, which motivates conservative absolute bounds tied to biopotential and acquisition physics [27, 77–79]. In I-CARE, aggressive heuristic cleaning would risk deleting the pathology under study, while physically anchored thresholds can be fixed before cohort-wide processing and audited directly during calibration. This design is therefore a domain requirement rather than a stylistic preference.

## 3.4 Feature Engineering

### 3.4.1 The 60-Second Window Rationale

A primary challenge in ICU EEG is its pronounced non-stationarity over hours to days: both the background pattern itself and the predictive value of specific features shift as sedation, temperature management, and injury evolution progress [39, 70, 73, 80]. To resolve this, the architecture adopts the **60-second window** as the basic quasi-stationary unit. This duration is selected to be long enough to capture meaningful spectral distributions and burst-suppression cycles, while remaining short enough to minimise the impact of state-transitions within a single epoch.

### 3.4.2 Hierarchical Patient-Level Aggregation

The reduction of the temporal window sequence to a stable patient-level biometric is governed by a *Hierarchical Aggregation* protocol. Because comatose EEG is highly non-stationary, simple temporal averaging of features is often mathematically destructive—for instance, averaging a high-voltage burst state and a low-voltage suppression state results in a mean value that represents a physiological state the brain never actually occupies. To resolve this, features are aggregated based on their neurophysiological typology:

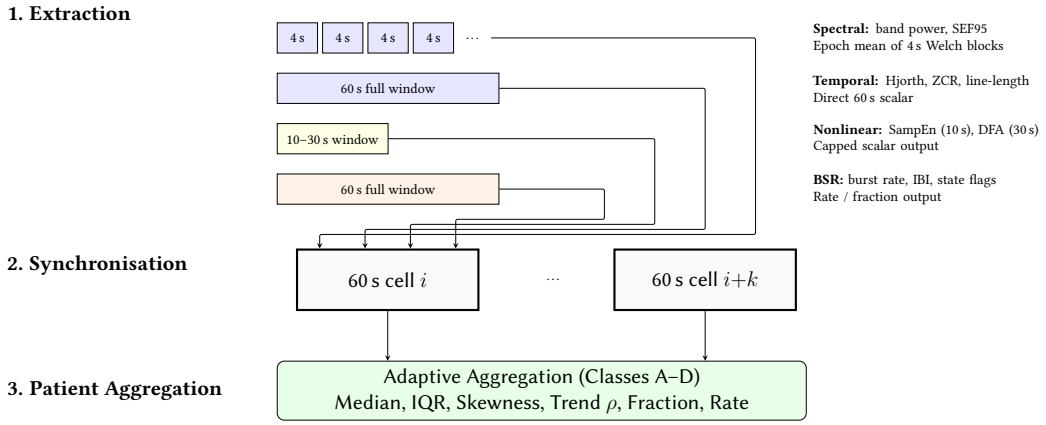



**Figure 3.4:** Multi-scale temporal extraction and synchronisation grid.

Although all features share the same 60-second synchronisation epoch (Level 2), the Level 1 effective analysis window differs by feature family. Temporal features (Hjorth parameters, ZCR, line-length) are computed over the full 60-second epoch and map 1:1 into the synchronisation cell. Spectral features use 4-second Welch sub-segments (`nperseg` = 1024 at 256 Hz), yielding 0.25 Hz frequency resolution; the sub-segment

PSDs are averaged into a single spectrum per epoch. Nonlinear complexity features (sample entropy, DFA exponent, Higuchi FD) are capped at the first 10–30 seconds of each epoch to keep per-channel computation below 1 second while preserving estimate validity (see Section A.4.4). Burst-suppression features operate on the full 60-second epoch with a 0.5-second minimum event duration (ACNS criteria).

- **Continuous:** High-stability spectral and aperiodic traits (band power, PAF, aperiodic exponent/offset, spectral entropy, SEF95, Hjorth parameters, wavelet energy) aggregated via robust central tendency (median, IQR).
- **Bimodal:** State-switching traits with bimodal window-level distributions (BSR, burst rate, burst amplitude/duration, IBI mean/CV, suppression depth) aggregated via mode-aware statistics to preserve the bimodal structure.
- **Binary:** Presence/absence flags for discrete pathological states (e.g. seizure detected, burst-suppression present, sigma-spindle flag) aggregated as fraction of windows positive.
- **Count:** Event-based features (e.g. spike rate, burst count, sharp-transient rate) normalised to per-hour rates across the aggregation window.

ACNS-aligned quantitative proxies for clinical EEG patterns — spike activity, periodic discharge regularity, sharp-transient morphology, and high-frequency oscillations — are extracted by dedicated modules and assigned to the Binary or Count class as appropriate; Section A.4.3 provides the complete feature catalogue.

### 3.4.3 Aggregation Class Distribution

The distribution of aggregation classes across feature domains demonstrates the domain-specific nature of the statistical moments required to capture temporal variability. See Section A.5 for the complete classification catalogue.

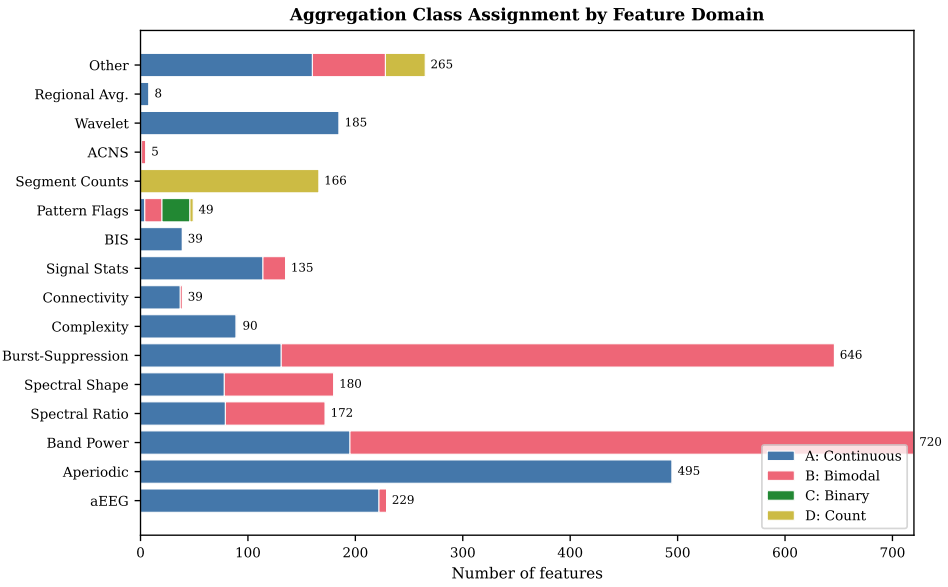



**Figure 3.5:** Distribution of the four aggregation classes across feature domains. The Burst-Suppression domain is not a single binary pathology flag; it aggregates channel-wise and global descriptors (BSR, IBI, burst duration/amplitude, suppression depth), which explains its large feature count.

## 3.5 Phase 1 Design: Feature-Based Machine Learning

Phase 1 provides the interpretable baseline. Before attempting higher-capacity sequence models, the study first tests whether an age-related signal is recoverable from established neurophysiological features at all.

The rationale for Phase 1 is three-fold:

1. **Interpretable Baseline:** By using handcrafted features (spectral power, connectivity, fractal complexity), this phase establishes a benchmark grounded in known comatose EEG patterns and follows prior EEG benchmarking work where classical models provide robust reference performance before higher-capacity models are trusted [67, 81].

2. **Feature-State Association:** This phase explicitly tests whether specific clinical patterns (e.g., alpha-coma, burst-suppression) drive the age-prediction, using SHAP-based feature-importance analysis to audit the model's decision logic and inform the severity validation check (H-7).

3. **Diagnostic Check:** By attempting the auxiliary task of sex classification on the same feature set, the information content of the extracted representation is tested

before stronger claims are made about age recovery.

Phase 1 employs **Profile B** (Bipolar derivation), which prioritises the rejection of cardiac volume conduction and non-stationary baseline drift—the primary confounds for frequency-domain feature extraction in the ICU.

### 3.5.1 Age-Bias Correction

All age models suffer from regression-to-the-mean, where younger patients are overpredicted and older patients underpredicted [82]. A mandatory post-hoc bias correction is enforced via linear residualisation:

$$\text{CN-BAG}_{\text{corrected},i} = \text{CN-BAG}_{\text{raw},i} - (\alpha + \beta \cdot a_i) \tag{3.3}$$

where $\alpha$ and $\beta$ are the intercept and slope derived from the training data. This ensures the metric purely represents physiological deviation rather than statistical artefact.

## 3.6 Phase 2 Design: End-to-End Deep Learning

Phase 2 tests whether end-to-end deep learning architectures can recover latent temporal dynamics and high-order cross-frequency relationships that are effectively invisible to window-level feature aggregation. In design terms, coma is treated as a pathologically recalibrated state rather than a lightly suppressed version of wakefulness, so the sequence model is asked to learn the residual structure that survives that recalibration. This is motivated by the fact that post-arrest EEG changes systematically over time under sedation, temperature management, and evolving injury, so time-since-arrest cannot be treated as a nuisance that is simply averaged away [70, 73]. In the final thesis configuration, Phase 1 aggregates handcrafted features over 60-second windows, whereas Phase 2 operates on overlapping 20-second waveform windows.

While Phase 1 provides the anchor, Phase 2 explores whether the age-related variance $\|f(a_i)\|$ is hidden in the *temporal transition logic* of the signal (e.g., the specific morphology of burst transitions or the non-linear coupling between bands). To achieve this, this phase employs **Profile C** as the baseline tensor path: a minimally filtered, common-average-referenced signal that preserves waveform continuity and does not force cardiac removal up front. The matched **Profile C-noecg** ablation adds targeted ECG-ICA to test whether removing the cardiac component materially changes downstream representation

learning; the full profile definitions are listed in Section A.1.1.

Phase 2 was therefore designed as a two-step process rather than a single fixed-network run. The initial architecture screen (EXP-200) compared four candidate backbones under the same Profile C waveform regime: `CNNBaseline` (pure convolutional ablation), `MultiTaskBrainAge` (CNN–BiLSTM–attention), `TCNMultiTask` (dilated temporal convolutions), and `EEGConformerMultiTask` (shallow convolution plus self-attention). Both the screen and the subsequent canonical run (EXP-201) were executed on the 286 development patients whose Profile C recordings met the Stage 1 hardware-fault threshold (out of 483), with a per-patient window cap of 50—a practical ceiling imposed by GPU memory capacity and SLURM wall-time limits on the HPC cluster. This comparison tested whether recurrent context, dilated convolutions, or transformer-style attention added measurable value over a shared CNN encoder on comatose EEG. The candidates were deliberately positioned across the data-requirement spectrum: global self-attention (`EEGConformerMultiTask`) achieves state-of-the-art performance on large EEG corpora but is known to underperform on smaller training sets—a limitation especially relevant at the $N = 604$ cohort scale [49]; `TCNMultiTask` offers causal dilated convolutions that extend the effective receptive field without recurrent bottlenecks; and `MultiTaskBrainAge` introduces bidirectional temporal context that has proved effective specifically on post-cardiac-arrest comatose EEG [83, 84]. Detailed layer specifications are provided in Section B.12, and the empirical comparison is summarised in Table B.16.

### 3.6.1 CNN-BiLSTM-Attention Architecture

This class of hybrid architecture—combining a convolutional encoder with a BiLSTM and an attention aggregator—has been applied to outcome prediction in post-cardiac-arrest comatose EEG [83] and to brain age estimation from non-wakefulness EEG [85]. The Phase 2 model (`MultiTaskBrainAge`) adapts this topology to ICU brain age prediction:

1. **Convolutional Encoder:** A series of 1D-CNN layers compress the continuous waveform into a sequence of latent representations, effectively learning a purpose-built feature extractor for comatose EEG.

2. **BiLSTM Sequence Layer:** Captures temporal dependencies within the 20-second waveform window, identifying patterns in local state transitions.

3. **Attention Aggregator:** Dynamically weights the temporal segments, allowing the model to focus on "neurophysiologically rich" periods while ignoring masked or suppressed segments.

Crucially, Phase 2 integrates the **Quality Mask** as a parallel input tensor, allowing the attention mechanism to "learn" the hardware failure states rather than blindly processing them.

### 3.6.2 Multi-Task Learning Rationale

To stabilise the representation, Phase 2 employs **Multi-Task Training** (**H-11**). By simultaneously training the model to predict age (regression) and sex (classification), the shared convolutional encoder is forced to learn "Universal Trait Representations" that are robust to state-dependent fluctuations. If the model can recover the biological sex of the comatose patient, it provides a powerful validation that the latent space contains premorbid biological information.

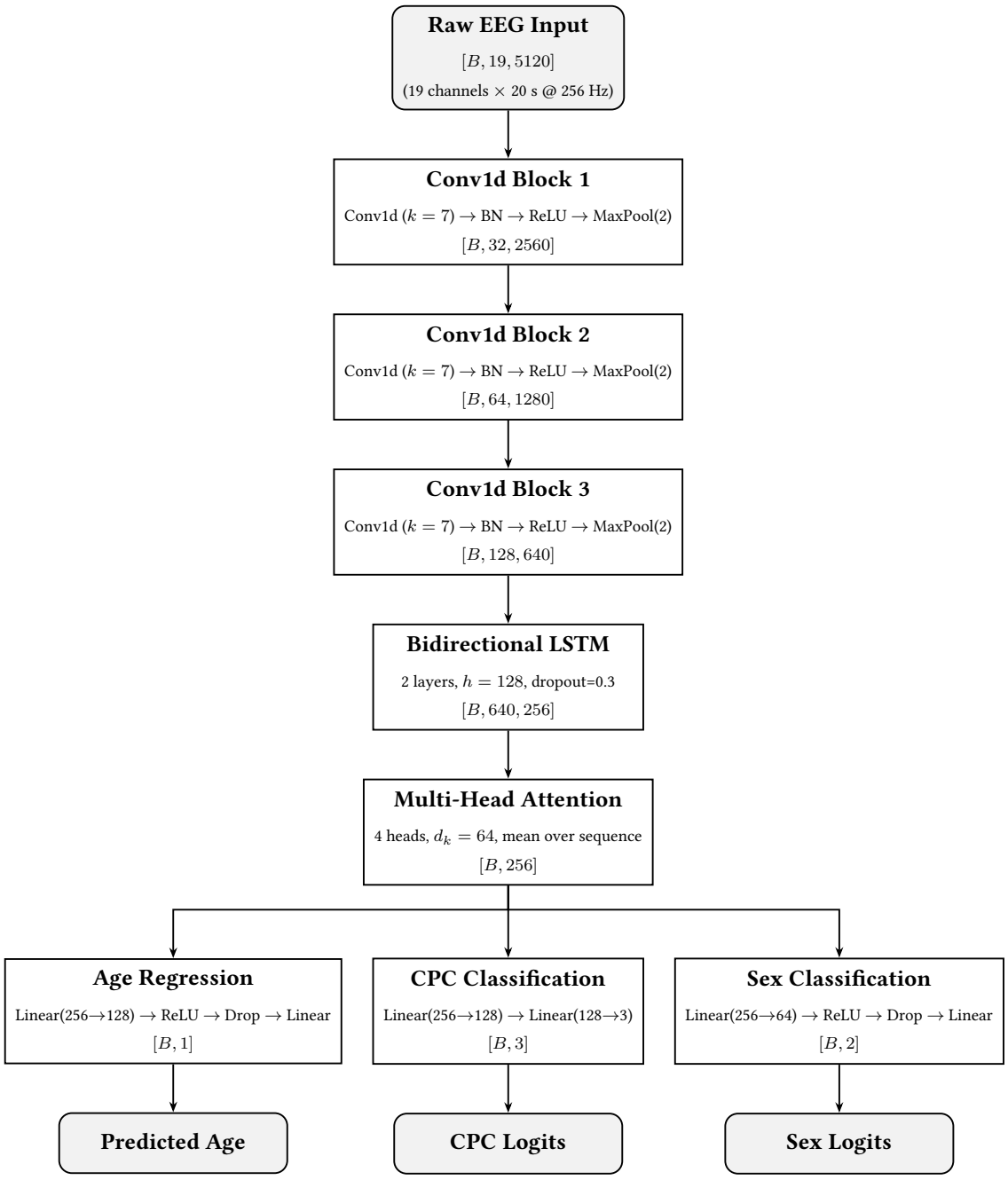



**Figure 3.6:** Phase 2 CNN-BiLSTM-Attention architecture (`MultiTaskBrainAge`). Profile C tensors are compressed via 1D-CNN filters, sequenced by a BiLSTM, and aggregated via attention to predict age and sex simultaneously.

## 3.7 Dataset and Infrastructure

### 3.7.1 The I-CARE Consortium Dataset

The analysis utilises the I-CARE database, a multi-centre corpus comprising $\approx 1.7\ TiB$ of continuous EEG from seven academic hospitals [61]. The inclusion criteria require patients to be $\geq$18 years of age, have suffered a return of spontaneous circulation (ROSC) following cardiac arrest, and remain in a comatose state (GCS-M $\leq$3) during the initial recording. Of the 1,020 patients in the archive, 607 were publicly released via the PhysioNet Challenge 2023 subset; after excluding three patients with missing demographic records, the final analysable cohort consists of $N = 604$ patients with a median recording duration of 46.1 hours.

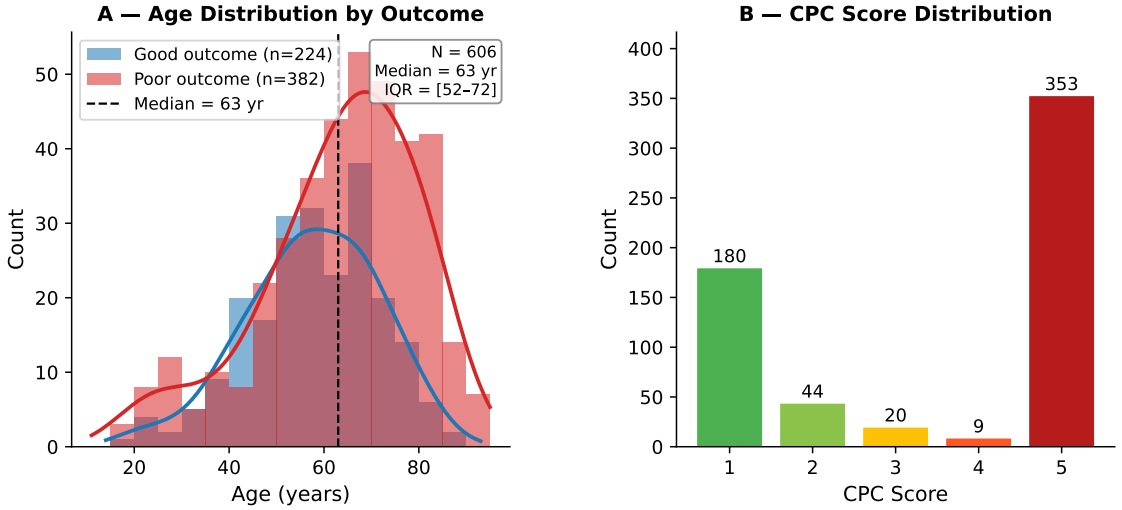



**Figure 3.7:** Demographic distribution of the I-CARE cohort ($N = 604$). Panel A illustrates the age distribution (Median 63 y). Panel B shows the distribution of Cerebral Performance Category (CPC) outcomes.

A primary challenge of the I-CARE cohort is the **extreme variance in recording lengths**, which range from 0.4 h to 345 h. This heterogeneity, combined with sampling rate variations across the seven participating centres, necessitates the robust Phase 0 preprocessing infrastructure described in Section 3.3. Beyond recording-length heterogeneity, the Targeted Temperature Management (TTM) protocol constitutes a structured clinical confound that introduces systematic temporal drift in the signal: EEG oscillation frequencies and amplitudes shift predictably across the cooling, rewarming, and normothermic phases, so time-post-arrest is treated as a mandatory trajectory covariate and patient-level cross-validation splits are enforced throughout [73].

The analysis window is bounded to the **first 36 hours** post-arrest. This reflects a pragmatic compromise: processing full-duration recordings was not feasible within the available computational and storage constraints. Anchoring to the earliest available signal provides a shared temporal reference point and avoids systematic availability bias towards patients with longer ICU stays. Patients whose recordings end before 36 hours contribute all available signal.

Dedicated ECG channels are available in 68.5% of patients (416 of 607); for this subset, the co-recorded cardiac trace enables EECI computation (Section 3.3.8) as a multi-level cardiac contamination diagnostic—quantifying cardiac coupling through temporal cross-correlation, spectral coherence, and phase-locking rather than simple frequency-band subtraction. For all patients, regardless of ECG availability, bipolar spatial derivation (Section 3.3.3) provides passive cardiac-artefact attenuation without requiring a reference cardiac channel; the two approaches are therefore complementary.

### 3.7.2 Missingness and Specparam Failures

A salient challenge in the feature engineering is the handling of spectral parameterisation ($1/f$ slope and offset). The analysis identifies a *Not Missing At Random (NMAR)* phenomenon: the `specparam` algorithm frequently fails on highly suppressed or isoelectric windows. Standard median imputation is rejected for these failures. Instead, because failure to converge is itself a biomarker of severe injury, the pipeline propagates a binary *`specparam_failed`* flag. This allows downstream models (XGBoost) to use the failure state as a high-importance pathological indicator.

A related consideration is feature collinearity: EEG channels share common sources via volume conduction, so spectral, connectivity, and temporal features derived from nearby electrode pairs carry substantial intercorrelation; the effective degrees of freedom in the 1,578-feature matrix are substantially smaller than the nominal count. ElasticNet (L1+L2), Random Forest, and XGBoost all apply regularisation that suppresses redundant predictors natively, so collinearity does not threaten model validity, but feature-importance rankings should be interpreted at the level of functional groups rather than individual features.

## 3.8 Scientific Integrity and Quality Assurance

The methodological integrity of this study is governed by a statistical analysis plan defined prior to final analysis, together with a validation framework designed to mitigate Type I error inflation and control for clinical confounds.

### 3.8.1 Quality Gates

The design integrates diagnostic quality gates throughout the preprocessing pipeline. These gates do not function as patient-exclusion criteria; instead they verify data completeness, monitor drift, and expose transformation failures before downstream modelling proceeds. If a severe error is detected, the affected epoch is flagged or propagated as low-reliability rather than the patient being removed. Dashboard-based telemetry and the broader reproducibility framework are described in Sections C.1 and C.4.

**Distinguishing CN-BAG from a Severity Proxy.** The CN-BAG does not claim to isolate a pure biological age signal from the compound of age, injury, and pharmacology; it is understood as a population-relative summary statistic that may partially overlap with severity indices. To characterise this overlap empirically, the study computes the *Jaccard similarity at top-20 features (Jaccard@20)* between the CN-BAG model and a clinical outcome regressor. Low overlap would provide evidence that the CN-BAG captures variance beyond what severity alone explains; high overlap would be an interpretability constraint to report transparently.

### 3.8.2 Data and Information Leakage Prevention

Windowed multi-recording EEG introduces leakage pathways that extend beyond the standard train/test boundary. A model that exploits any of these pathways can report plausible-looking error metrics while having learned nothing generalisable about biological age. This study addresses four nested threat levels, each closed by a specific design choice, and acknowledges two residual dependencies that are mitigated but not fully eliminated.

**(1) Subject-level leakage.** The most fundamental risk is that the same patient's windows appear in both training and evaluation, allowing the model to memorise idiosyncratic patient signatures. This is blocked by grouping cross-validation folds strictly on patient identity throughout the development set, and by the sealed test set (Section 3.2.2),

which is withheld from all tuning.

**(2) Window-level pseudo-replication.** Each patient contributes hundreds of temporally autocorrelated windows. Randomly allocating those windows across train and validation inflates apparent performance because the model has effectively seen the patient before—just at a different timestamp. The intra-patient temporal holdout (Section 3.2.2) closes this cross-boundary gap: training windows are drawn exclusively from the chronological first $80\%$ of each patient's recording; the final $20\%$ is reserved as a within-patient validation segment inaccessible during training.

**(3) Artefact-as-proxy leakage.** Hardware fault patterns and electrode noise profiles may correlate with patient age, recording site, or recording duration in ways opaque to standard preprocessing. A model unaware of quality states may exploit these correlates as shortcuts to apparent trait recovery. The quality mask (Section 3.3.6) is passed as an explicit input tensor to the Phase 2 architecture, so the attention mechanism represents artefact states transparently rather than absorbing them silently into the learned representations.

**(4) Confound-as-shortcut leakage.** Coma severity, sedation depth, and TTM phase all co-vary with EEG morphology and, indirectly, with age. A model that learns severity rather than biological age would pass standard error checks while providing no genuine prognostic signal. The multi-task sex classification control (Sections 3.2.2 and 3.6) provides a continuous diagnostic check: sex is entirely independent of coma severity, so a model that collapses onto a severity shortcut simultaneously loses sex discriminability. The Jaccard@20 feature-overlap analysis (Section 3.8.1) supplies a complementary quantitative check, comparing CN-BAG feature importance against clinical outcome regressors.

**Residual dependencies.** Two signal-structure dependencies are mitigated but not fully eliminated. *Temporal autocorrelation within the training segment:* EEG windows drawn from the same patient within the training pool are not independent; slowly varying state trajectories (evolving injury, TTM rewarming, sedation washout) mean the model is exposed to non-i.i.d. data even after the temporal holdout is applied. Patient-level aggregation at inference (median age across windows; majority-vote sex) partially absorbs within-patient variance, but does not correct for the inflated effective sample count during training. *Channel cross-dependence:* the 19 electrodes share a common reference and sample overlapping cortical volumes, so the 19 feature vectors or waveform channels per window are spatially correlated rather than independent observations. This does not constitute leakage in the strict sense—no label information crosses the train/test

boundary—but it reduces the effective information per window and means that confidence intervals derived from window counts overstate the true sample size. Both limitations are treated as acknowledged design constraints rather than resolved flaws; their practical impact is bounded by the patient-level evaluation metric reported throughout.

Together, the four active defences ensure that any observed model performance cannot be attributed to the leakage pathways endemic to windowed ICU EEG analysis, while the residual dependencies are transparently bounded.

### 3.8.3 Validation Strategy and Statistical Plan

To maintain narrative focus, the detailed mathematical formulations for these controls are documented in the project appendices:

- **Three-Tier Testing Strategy:** Every processing stage is governed by automated unit tests, integration gates, and a final visual audit of model outliers to ensure results are grounded in neurophysiology rather than artefactual shortcuts.
- **Hypothesis Testing:** The high-level Research Objectives (RO-1–3) are decomposed into eleven testable hypotheses. All eleven hypotheses (H-1 to H-11) were specified and documented before the experimental analysis was conducted. The sequentially rejective Bonferroni-Holm procedure used to control the family-wise error rate across the two primary families (F1: Outcome Association; F2: Trait Recovery), together with the within-analysis test-selection rationale, is detailed in Appendix Appendix E. Grouped cross-validation, patient-level outcome models, and empirical null evaluation follow prior EEG outcome and decoding literature that emphasises leakage control and non-parametric significance testing [73–75, 86]. The hypothesis-to-experiment crosswalk is given in Table B.2, and the order of H-1 to H-11 is fixed across the thesis.
- **Bias and Null-Model Controls:** The protocols for regression-to-mean correction and the permutation-based MAE floors used to validate feasibility (H-1) are documented in Appendix E.
- **Confound and Severity Validation:** The WLST Dual Analysis tracks and "Severity Shield" validation protocols are detailed in Appendix B.

At the development-set scale ($N = 483$), all primary hypotheses exceed 85% power; the permutation-based null-MAE floor for H-1 is $12.54$ y, and the pre-specified success

criterion of MAE $< 12$ y sits conservatively below this threshold to provide a margin above sampling variability in the null estimate. The sex-stratified hypothesis (H-10) is designated exploratory owing to reduced power ($57.6\%$ at $N \approx 150$). Full derivations are in Section E.3. A design limitation of the modelling approach is that all models treat the cohort as a single pooled group rather than stratifying by clinical subgroup (TTM target temperature 33°C vs. 36°C, OHCA vs. IHCA): splitting by these subgroups would reduce per-fold training sets to fewer than 80 patients for the minority classes, collapsing cross-validation reliability; subgroup effects are therefore examined as covariates in post-hoc regression rather than as separate modelling tracks. The transition from the scientific *Why* of this chapter to the experimental *How* of Chapter 4 is governed by an operationalisation roadmap where research objectives are formalised into testable hypotheses, and those in turn are operationalised as interlinked and purpose-driven experiments. This iterative, goal-oriented approach is accompanied by a continuous bottom-up approach, where enhancement in domain understanding, technological hurdles, and new findings about the development dataset are structurally incorporated into the design and refinement of upstream experiments. This refinement includes research objectives but occurs strictly prior to experimental Phases 1 and 2.

## 3.9 Project Governance and Management

The project management and execution of this study is governed by a comprehensive project methodology designed to ensure that scientific standards are met and systematically integrated in the methodological and experimental design. At a high level, scientific decisions, experiment tracking, and implementation outputs were kept explicitly linked so that every reported result could be traced back to a fixed configuration and commit state.

### 3.9.1 Project Organisation

The project was partitioned into functional modules spanning governance, preprocessing, feature extraction, modelling, evaluation, and infrastructure. That modular structure was not just organisational overhead: it was the practical mechanism by which a decentralised, AI-assisted project of this size remained inspectable and reproducible. These modules interact with the data pipeline through a dedicated *Control Plane* (visualised in Figure 3.1), comprising quality assurance, exploratory data analysis, infrastructure, scientific gov-

ernance, and evaluation functions that operate continuously across all pipeline stages. The full module taxonomy, dashboard views, and governance rules are documented in Sections C.1, C.1.2 and C.4.

### 3.9.2 Computational Reproducibility and Determinism

To ensure that all experimental findings are strictly reproducible, technical safeguards are integrated into the execution environment:

- **Global Seed Standardisation:** Random number generators across NumPy, PyTorch, and scikit-learn are initialised with a fixed global seed to guarantee deterministic data partitioning and network initialisations.
- **Git Hash Tracking:** Every HPC job launched via the runner script automatically captures and logs the active Git commit hash. This immutably links every result, generated model weight, metric, and artefact back to the exact codebase state that produced it.
- **CUDA Determinism:** For Phase 2 deep learning pipelines, PyTorch is configured to enforce deterministic algorithms (e.g., `torch.backends.cudnn.deterministic = True`). This ensures that GPU-accelerated operations yield reproducible gradients and weights despite non-deterministic hardware threading, accepting a computational overhead in exchange for absolute scientific traceability.

### 3.9.3 AI-Assisted Research Workflow

The workflows of this project were organised as an iterative research loop: define research objectives, derive hypotheses, operationalise them as experiments, build the necessary tooling, and refine the design against development-set diagnostics before final evaluation.

Generative AI tools were integrated into the research process under continuous human oversight. Retrieval-augmented literature workflows and tool-adaptations were developed to manage and fully utilise the corpus of selected works for background literature across the clinical and neurophysiological domains and data science methodology. Decentralised project dashboards were not incidental conveniences: they were essential work tools designed for the specific task of managing a large, decentralised data-science pipeline whose code, notes, experiments, and thesis text had to remain synchronised. In the case

of the modular task dashboard, the structure was specifically tailored to AI-requirements as a modular, serverless JSON-file system. The researcher remained the system architect and final scientific authority throughout. The detailed tool-to-task mapping, governance boundaries, and dashboard descriptions are documented in Sections C.4 and C.5.

This chapter has specified the scientific design: what is done, why, and under what assumptions. Chapter 4 describes how this design is operationalised — the iterative validation, quality assurance, and engineering solutions that bridge the analysis plan and the results.

# Chapter 4

# Experiments

Chapter 3 defines the study design, modelling logic, and validation framework. This chapter documents how that design was executed against the I-CARE corpus: how the cohort was partitioned and the gated execution framework established, the verification chain that locked preprocessing thresholds, the multi-profile preprocessing runs, the feature-extraction workflow, the two modelling tracks, and the reproducibility controls applied throughout.

The chapter is ordered to mirror the data flow through the pipeline. The experimental architecture and data partitioning (Section 4.1) are presented first because the development–test split is a logical prerequisite for every downstream step: the verification chain (Section 4.2) was executed entirely within the development set, and the preprocessing profiles (Section 4.3) were locked before any modelling began. The feature-engineering workflow (Section 4.4) consumed those fixed profile outputs, and the two modelling tracks (Sections 4.5 and 4.6) then ingested the feature matrices and waveform tensors produced upstream. The dataset and infrastructure overview (Section 4.7) and the continuous quality assurance and workflow controls (Sections 4.8 and 4.9) follow. Technical scaffolding that would have overburdened the main execution narrative is consolidated in the appendix and cross-referenced throughout.

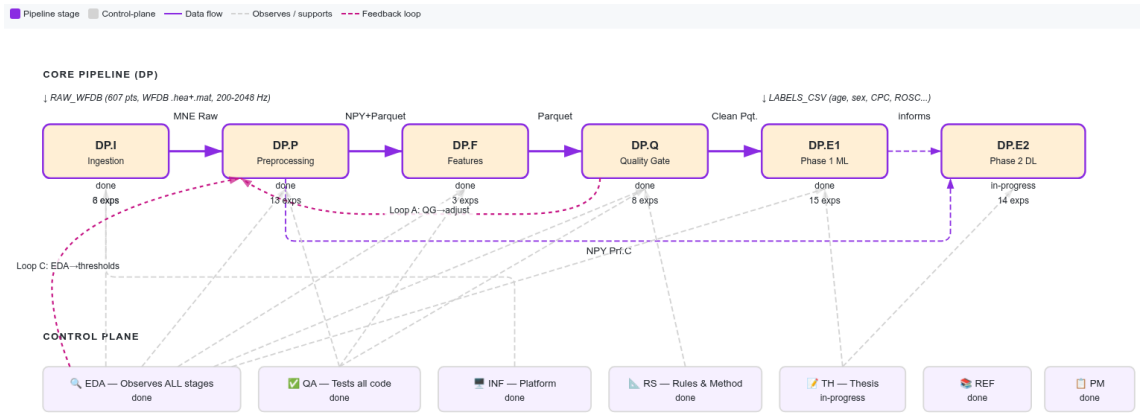



**Figure 4.1:** End-to-end experimental pipeline flow, from the raw I-CARE archive through the gated quality infrastructure, multi-profile preprocessing, feature engineering, and dual-track modelling (Phase 1 / Phase 2), to final locked-test evaluation.

## 4.1 Experimental Architecture and Partitioning

Figure 4.1 illustrates the end-to-end data flow. Raw WFDB recordings from the I-CARE archive enter the ingestion stage alongside the accompanying labels (age, sex, CPC score, and ROSC time). From there the signal passes through preprocessing, feature extraction, and a quality gate before reaching the two modelling tracks; Profile C waveform tensors bypass the feature layer and flow directly into the deep-learning track. Two feedback loops are visible in the figure: empirical characterisation runs on the development cohort informed the artefact-rejection thresholds before full-cohort preprocessing began (Loop C, Section 4.2), and quality-gate outcomes fed adjustments back into the preprocessing configuration (Loop A). Phase 1 results additionally informed the design of the Phase 2 deep-learning architecture before Phase 2 training commenced. A control plane—comprising the EDA module, continuous QA, the infrastructure layer, and the rules-and-method record—observed and supported all pipeline stages throughout.

The development–test boundary was drawn with sex stratification (`seed=42`), preserving the cohort's male-predominant composition (approximately 69%; a characteristic of post-cardiac-arrest epidemiology) proportionally across both partitions. Age and clinical outcome distributions were additionally verified for balance across the boundary; no significant distributional shifts were detected (Table 4.1). The Phase 1 sub-split (386 training / 97 holdout patients) was drawn with sex stratification likewise (`seed=43`), with age and outcome balance verified before model training commenced.

Threshold validation (Section 4.2) was performed exclusively within the 483-patient development cohort. Computational execution was orchestrated via SLURM batch jobs

**Table 4.1:** Detailed Demographic Balance (Development vs. Locked Test). No significant distributional shifts exist between the training and evaluation partitions.

| **Feature** | **Development** ($n = 483$) | **Locked Test** ($n = 121$) | **p-value** |
|---|---|---|---|
| Age (Median [IQR]) | 63.0 [52.0, 72.0] | 63.0 [53.0, 72.0] | 0.96 |
| Sex (Male %) | 333 (68.9%) | 84 (69.4%) | 1.00 |
| OHCA (%) | 353 (78.1%) | 89 (78.1%) | 1.00 |
| Shockable Rhythm (%) | 237 (51.5%) | 60 (52.6%) | 0.89 |
| CPC 1–2 (Good) | 181 (37.4%) | 44 (36.4%) | 0.88 |

on the GWDG KISSKI cluster (Section 4.9).

## 4.2 Engineering Execution and Validation

An *empirical verification chain* was executed prior to full-cohort processing to lock the artefact-rejection thresholds before the main runs. A 20-patient stratified probe first confirmed that the quality-mask bitmasks captured genuine hardware faults without over-flagging pathological bursts. The main chain then progressed through three sequenced experiments:

1. `EXP-001-A`[1] *(Profile A Baseline):* The full 483-patient development cohort was processed with only anti-aliasing resampling and extreme artefact masking, generating the raw empirical EEG distributions of the comatose state.

2. `EXP-012` *(Artefact Quantification):* The Profile A distributions were profiled to quantify the prevalence of line noise, muscular artefacts, and cardiac volume conduction, providing the empirical basis for threshold anchoring.

3. `EXP-306` *(Quality Gate Lock):* Based on the distributions from `EXP-012`, the final rejection thresholds were frozen in a version-controlled configuration matrix. Profiles B–E were only permitted to execute after this gate cleared.

### 4.2.1 Threshold Validation via CDF Analysis

No published consensus defines amplitude rejection thresholds for ICU EEG, and values reported in the literature span a wide range depending on acquisition system and patient population [77, 78]. For a brain-age study this uncertainty is particularly consequential: it is not known *a priori* which spectral components carry age-relevant information in the comatose state, so removing any frequency band or morphological feature risks discarding signal the model needs. The aperiodic slope—the most robust EEG brain-age predictor—is sensitive to aggressive filtering and step-removal [27]; burst-suppression transients carry age-relevant amplitude and entropy information but are routinely flagged as artefacts by standard detectors [87]; and cardiac volume conduction occupies the same 1–3 Hz delta band as coma-related neural slowing, making aggressive cardiac removal inseparable from removal of genuine neural signal. The full methodological rationale is developed in Sections 3.3.4 and 3.3.10; the consequence here is that the amplitude

[1] All experiments executed in this study are listed in Table B.3 (Appendix B), grouped by the eight pipeline stages from infrastructure validation through outcome association. Full results, per-fold breakdowns, and hypothesis verdicts are reported in Appendix B and Chapter 5.

limits ($500\,\mu$V saturation, $0.01\,\mu$V flatline, $200\,\mu$V/sample jump) were defined *a priori* from hardware physics rather than fitted to the I-CARE data.

Two experiments operationalised this conservative stance. `EXP-001-A` processed the full 483-patient development cohort *without* the quality mask—preventing circular bias—to extract a raw signal atlas: Welch PSD band powers, aperiodic exponent, amplitude histograms at 200-bin resolution, and per-channel jump distributions. `EXP-012` then computed a *Cleaning Impact Score* (CIS) for each candidate preprocessing step, combining a decontamination term (reduction in cardiac contamination measured via the EECI) and a preservation term (retention of brain-age features including the aperiodic slope, band-power ratios, and Hjorth parameters), weighted equally; steps with CIS $\leq 0$ were classified as harmful and excluded from the primary profiles. A cumulative distribution function (CDF) analysis—mapping the empirical quantile function of amplitude and jump magnitude across all development-set observations—then verified where the pre-set hardware-physics limits fall within the measured distributions. The EECI cardiac contamination metric (Section 3.3.8) was used as the decontamination component of the CIS.

This empirical verification is distinct from formal external validation against expert-annotated artefact labels, which was not performed; no independently labelled artefact corpus for the I-CARE EEG is publicly available, and this constitutes a recognised limitation (Section B.2.4). The formal bit definitions and threshold logic are provided in Section A.2.

1. **Gap-Grounded Limits:** The analysis confirmed that the $500\,\mu$V saturation and $0.01\,\mu$V flatline thresholds sit within natural mathematical gaps in the comatose EEG distribution, preserving burst-suppression while masking amplifier faults.

2. **Jump threshold ($200\,\mu$V/sample):** Unlike the flatline limit, the jump distribution is continuous; the $200\,\mu$V/sample bound acts as a hard-physics ceiling at the maximum physiological EEG slew rate rather than a distributional separator. See Section A.2 for the full two-stage quality mask definition.

**Calibration versus validation.** Two distinct quality-control concepts operate at different levels of this pipeline and should not be conflated. *Validation* denotes the enforcement of fixed, hardware-physics-anchored thresholds (the quality mask): these limits are not fitted to the data but are applied post hoc to tag technically invalid samples, and they

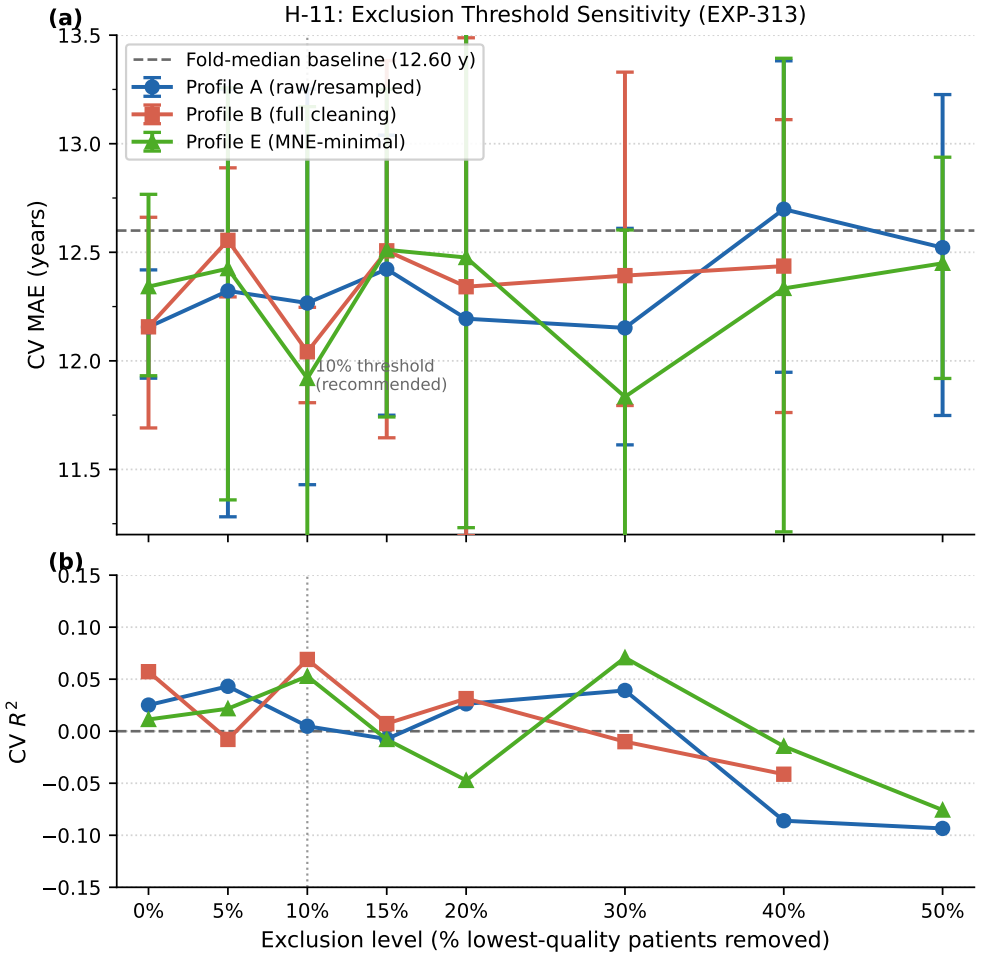



**Figure 4.2:** Quality-mask threshold sensitivity sweep (EXP-313, hypothesis H-9). Each data point shows the balanced accuracy of the held-out logistic-regression outcome model as the Stage 1 amplitude rejection threshold is varied. The dotted vertical line marks the locked hardware-physics threshold (500 $\mu$V). Model performance is stable across the empirically observed range, confirming that conservative threshold placement does not inflate or suppress the outcome-prediction signal.

were confirmed against the observed distributions via the CDF analysis above. *Calibration* denotes the data-driven locking of preprocessing parameters that *are* fitted to the cohort: `EXP-306` ran a majority vote across the 483-patient development set to lock the notch-filter type (IIR, 463/483 votes), the step-removal algorithm (v2, 388/483 votes), and the per-archetype ICA threshold (0.90), writing the final settings to a version-controlled `adaptation.json`. Profile-specific runs were only permitted after this calibration gate cleared.

### 4.2.2 Profile-Specific Parameter Locking

Several preprocessing parameters were empirically determined during the verification chain rather than adopted from healthy-EEG defaults:

- *IIR vs. FIR notch filter:* A patient-level vote across the 483-patient development set (463/483 favouring IIR Butterworth) confirmed the filter-type selection for the prominence-gated notch (Section A.1.2). The sharper IIR roll-off is better suited to comatose EEG's lower amplitude variability, making it less prone to ringing artefacts than in awake recordings.

- *Profile C-noecg ECG-ICA threshold:* The 0.90 ICA–ECG correlation threshold was set conservatively. The CIS analysis (`EXP-012`) confirmed that reducing this threshold increases decontamination but drives the preservation term below acceptable levels: lower thresholds remove independent components whose spectral mass overlaps with the 1–3 Hz delta band, the primary carrier of coma-state neural activity and a key input to aperiodic-slope estimation. Because it is not known which subcomponents of the delta band carry age-relevant information, the threshold was fixed at the point where decontamination is achieved without measurable delta-band distortion. Profile C-noecg provides a cortical-only baseline by removing cardiac volume conduction via ECG-ICA, enabling a direct comparison against the retained-ECG Profile C. This comparison is scientifically necessary because cardiac volume conduction is itself age-predictive — ICA-based cardiac removal has been shown to reduce age-prediction $R^2$ by $-0.046$ [27] — indicating that a model trained on unfiltered EEG may partly track cardiac rather than cortical ageing.

- *Profile D RANSAC failure:* RANSAC bad-channel detection was abandoned after threshold verification confirmed that inter-channel correlation in comatose EEG (median $r = 0.46$, IQR: 0.37–0.58) falls systematically below the default $r <$

0.4 flagging threshold, causing 25% of channels to be misclassified as defective. This empirical failure is reported as a negative finding (Section 4.3.1) rather than discarded silently.

## 4.3 Preprocessing Implementation

The raw I-CARE signal (Figure 4.3) enters a gated masking layer (Figure 4.4) that excludes hardware-fault segments before profile-specific spectral cleaning is applied; the resulting before/after transformation and the per-stage signal decomposition for Profile C (the primary Phase 2 input) are shown in Figures 4.5 and 4.6 respectively; the equivalent stepwise transformations for Profile A and Profile B are provided in Figures A.3 and A.4 in Appendix A.

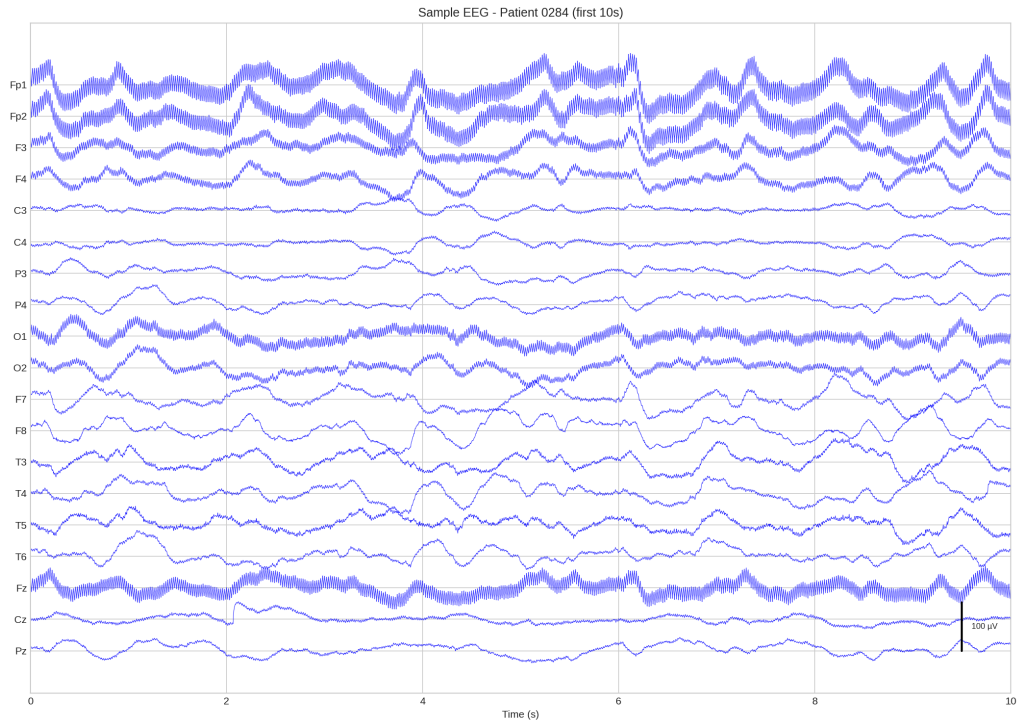



**Figure 4.3:** Raw EEG montage showing the standard 10-20 electrode configuration used across the I-CARE consortium sites. This raw signal serves as the input to the Stage 1 quality masking layer.

The preprocessing system was developed iteratively rather than as a monolithic workflow. An early monolithic prototype collapsed all five profiles into a sequential single-process loop, which proved computationally intractable and prevented fault isolation between profiles. The modular redesign separated loading, masking, and profile-

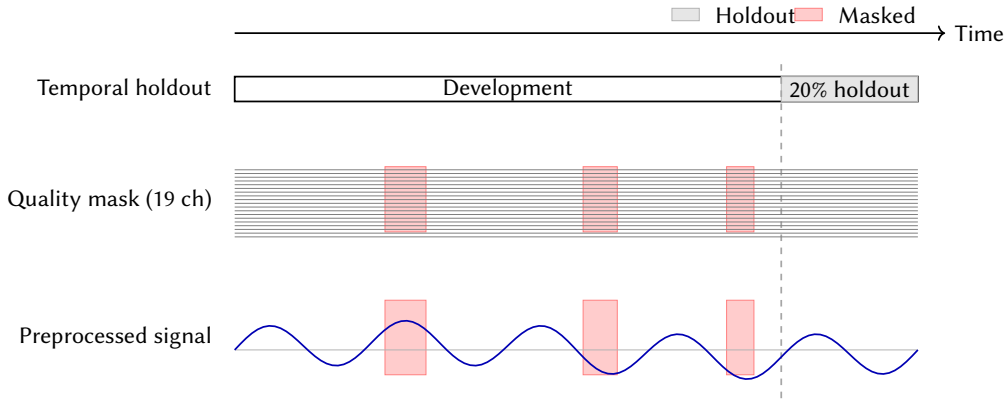



**Figure 4.4:** Temporal holdout and quality mask interaction. The dashed line marks the 20% temporal holdout boundary within each recording. Red bands indicate Stage 1 hardware-fault segments; a window is excluded from feature computation once the per-channel usable fraction falls below 50%, and Phase 2 attention ignores masked time steps. Full per-channel and global threshold definitions are given in Section A.2.

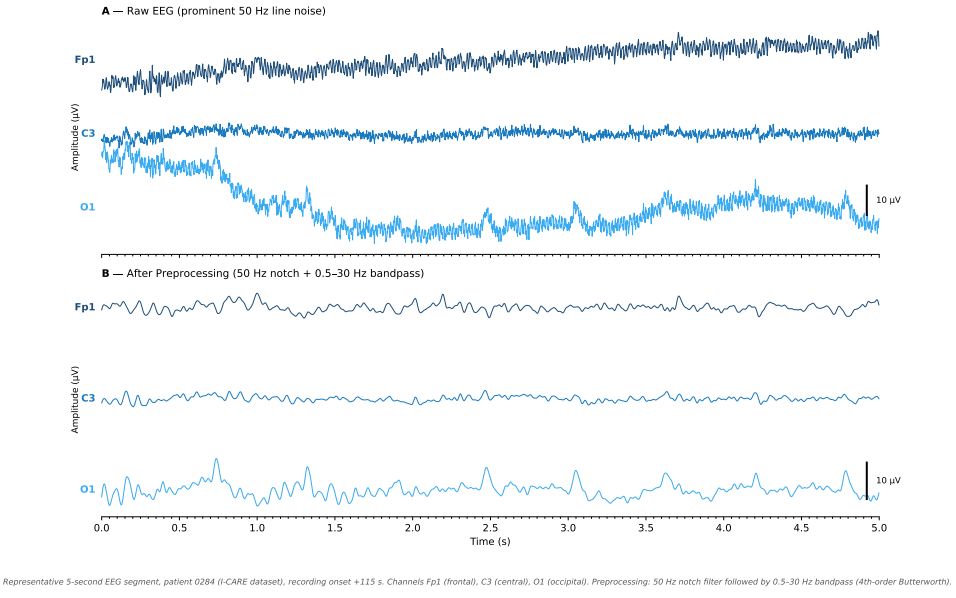
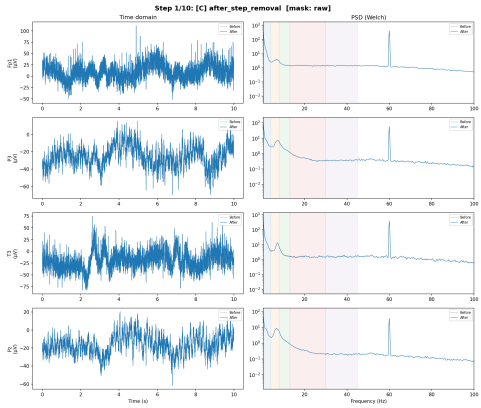
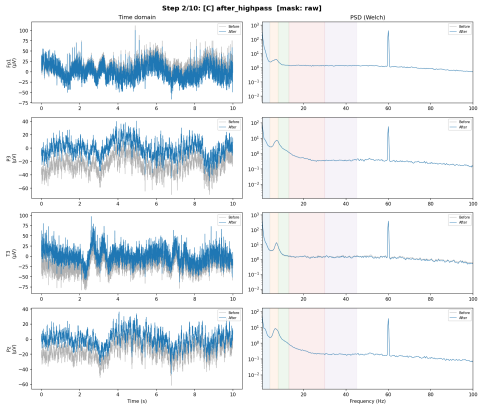


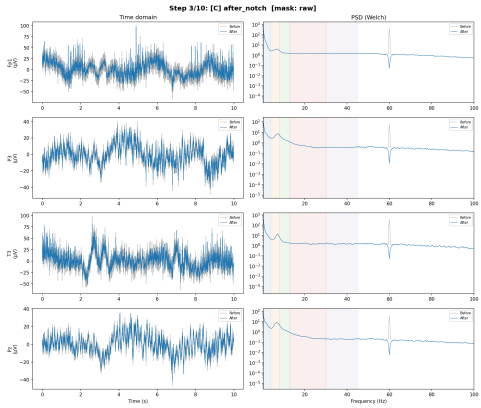

**Figure 4.5:** Visualisation of the preprocessing transformation from raw waveform to cleaned signal (Profile B). The design prioritises the suppression of utility noise and cardiac volume conduction while preserving the underlying neurophysiological delta power.



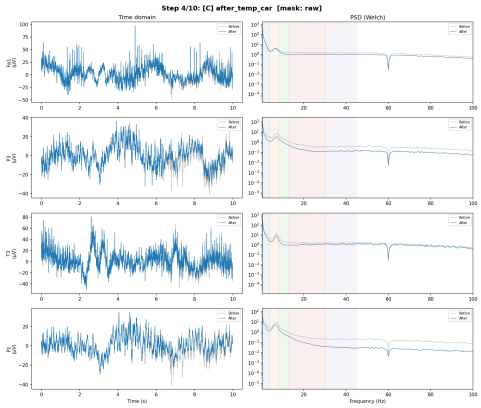

**(a)** After step-artefact removal



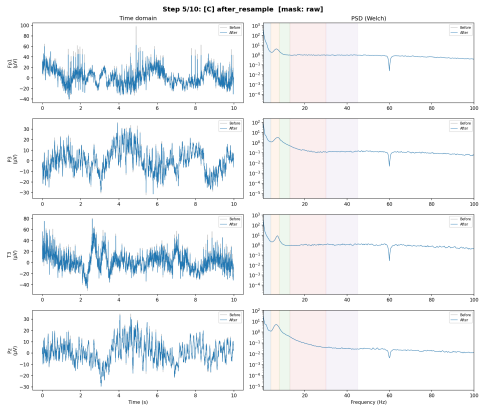

**(b)** After 0.5 Hz high-pass filter



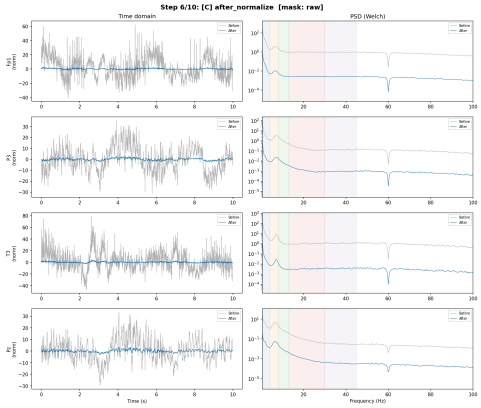

**(c)** After 50/60 Hz IIR notch




**(d)** After temporal CAR




**(e)** After resample to 256 Hz




**(f)** After patient-level robust normalisation (final output)

**Figure 4.6:** Profile C stepwise signal transformation for patient 0286. Each panel shows a representative 10-second EEG segment (temporal CAR re-reference) at successive pipeline stages. The high-pass filter removes low-frequency drift; the IIR notch suppresses mains interference; temporal CAR reduces correlated noise across channels; resampling standardises the sample rate to 256 Hz; and the final patient-level robust normalisation centres each channel for Phase 2 deep-learning input.

specific filtering into independent, composable modules with explicit contracts, allowing individual steps to be replaced or validated without reprocessing the full cohort. This modularity was essential for the iterative verification chain described above and later enabled the monitoring and dashboard layer to expose failures at the correct stage rather than only after full runs had completed.

Feature extraction was made fault-tolerant with explicit NaN-guard logic: structural missingness was preserved as a first-class signal rather than silently imputed away. A raw-archive audit surfaced nine patients with raw `NaN` values and one dead channel spanning 921,600 consecutive null samples; these were handled via dual-stream sanitisation (median-fill for metrics, interpolation for filtering) and logged as hard failures before the verification cycle was frozen.

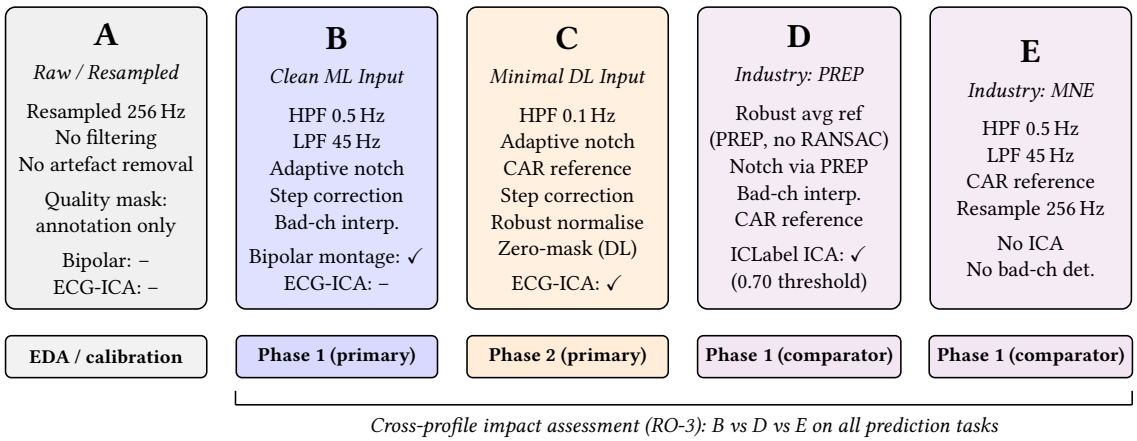



**Figure 4.7:** Stepwise preprocessing routing across Profiles A–E. The quality mask is applied uniformly at Stage 1 (Phase 0), after which each profile branches into its specific spectral cleaning, referencing, and normalisation steps. Profile D is shown as a failed comparator.

### 4.3.1 Multi-Profile Execution

The pipeline followed a strictly sequential multi-profile pass (A, B, C, E). For each patient, the raw WFDB waveform was loaded exactly once into RAM, and the profiles were executed in parallel sub-processes to bypass the Python Global Interpreter Lock (GIL).

**PREP/RANSAC incompatibility with coma EEG.** Profile D tested the industry-standard PREP/RANSAC bad-channel detector. Empirical profiling revealed a fundamental incompatibility: in comatose EEG, pathological cortical disconnection reduces inter-channel correlation to moderate levels (median $r = 0.46$, IQR 0.37–0.58; EXP-306),

well below the inter-channel agreement that PREP's flagging criteria assume. The detector cannot distinguish pathologically reduced neural connectivity from a defective electrode, causing it to flag functioning channels as bad and rendering the criterion unreliable for this cohort. RANSAC was disabled (`do_ransac=False`) for Profile D; bad-channel detection falls back to the absolute amplitude and variance thresholds of the Stage 1 quality mask. Profile D was therefore excluded from the primary analysis, confirming why the preprocessing design had to be validated iteratively rather than adopted from healthy-EEG defaults.

**Normalisation implementation.** Following the strategy specified in Section 3.3.9, normalisation is applied profile-specifically. Profile B outputs signals in raw $\mu$V without normalisation, since Phase 1 classical models operate on absolute-value features where amplitude magnitude carries neurophysiological meaning. Profiles C, D, and E apply *patient-level robust normalisation*: each patient's tensor is clipped to $\pm 10$ IQR and scaled by the patient's own median and IQR, preserving internal amplitude dynamics across the stay while preventing gradient saturation in the Phase 2 BiLSTM. Full per-profile settings are tabulated in Section A.1.1.

**Stage 1 hardware-fault exclusion.** The quality mask detects hardware-failure states in the time domain (flatlines, amplitude jumps $> 200$ $\mu$V/sample, saturation $> 500$ $\mu$V, electrode pops; thresholds and justification in Section A.2.1). The median fraction of 60-second windows passing all Stage 1 criteria was $0.669$ for Profile A (no spectral preprocessing) and $0.968$ for Profile E (full spectral cleaning stack; Wilcoxon rank-biserial $r = 0.997$, $p < 0.001$); the per-patient distribution is shown in Figure 4.8. Two caveats apply. First, Profile E's upstream step-removal reduces the amplitude transients that trigger Stage 1 criteria, so the difference partly reflects a processing-induced reduction in trigger conditions rather than a difference in underlying signal integrity. Second, Stage 1 detects hardware-level failure states only; windows that pass Stage 1 may still contain substantial physiological artefacts (cardiac volume conduction, EMG) annotated separately in the Stage 2 bitfield (Section A.2.3).

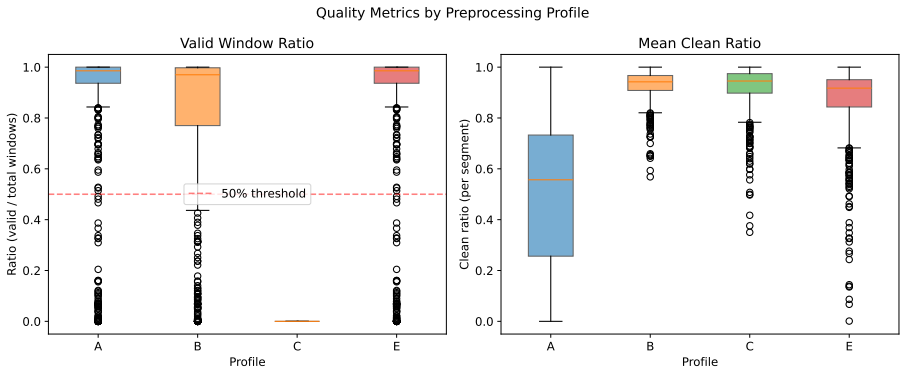



**Figure 4.8:** Fraction of 60-second windows passing all Stage 1 hardware-fault criteria per patient (Profile A: no spectral pre-processing; Profile E: full spectral cleaning stack). Profile E's lower exclusion rate reflects in part that upstream step-removal reduces the amplitude transients that trigger Stage 1 criteria; Stage 1 passage does not imply freedom from physiological artefacts.

## 4.4 Feature Engineering Workflow

Hierarchical aggregation was implemented as a distributed map-reduce operation. To manage temporal drift and keep the extractor fault-tolerant, the workflow used NaN-guarded feature handling, preserving structural missingness as explicit metadata instead of collapsing it into silent defaults. The extractor computed a rich set of spectral, temporal, and connectivity descriptors per 60-second window; these were then feature-adaptively aggregated across the recording to yield a high-dimensional patient-level feature matrix (Section 3.4.2). Representative visualisations of the extracted feature space on real ICU EEG, together with the full distribution, variability, and correlation analyses, are presented in Section A.7.

### 4.4.1 Exploratory Data Analysis and Temporal Feature Drift

The EDA characterised distributional properties of the full 483-patient development set. The two findings below determined the specific aggregation statistics employed within the feature-adaptive framework defined in Section 3.4.2:

**Kurtosis and pathological transients.** EDA revealed extreme kurtosis values (exceeding 20.0) in comatose segments, representing the physical reality of burst-suppression transients. This determined the use of percentile-based aggregation: simple temporal means would mathematically erase the high-voltage bursts that constitute the prognostically informative component of burst-suppression.

**Temporal feature drift.** Mapping the 1,578-feature matrix across the first 36 hours of the post-arrest monitoring window revealed pronounced systematic drift driven by clinical interventions. During the initial 24 hours, features are confounded by targeted temperature management (TTM) and maximum sedation, producing the metabolic frequency downshift described in Section 2.1. This confirmed that features must be aggregated into trajectories and variance summaries (IQR, skewness, slope) rather than static means.

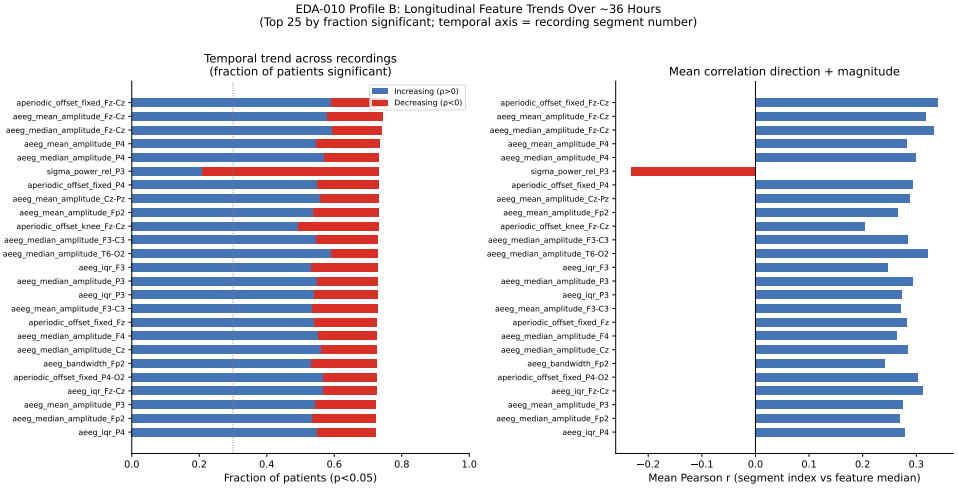



**Figure 4.9:** Temporal drift of key neurophysiological features across the first 36 hours of the post-arrest monitoring window (development set, $n = 483$). Significant monotonic temporal trends ($p < 0.05$, Spearman) were observed in 89.0% of features in more than 30% of patients, and in 68.4% of features in a strict majority (>50%) of patients, validating the necessity of trajectory-aware aggregation rather than static window-mean extraction.

An extended analysis — including feature-trajectory plots stratified by CPC outcome, KS-test discriminability scores, and the full EDA-010 suite — is presented in Section A.7.

### 4.4.2 Intra-Class Correlation and Feature Stability

The reliability of the extracted features was assessed via Intra-Class Correlation (ICC) across 60-second windows (Figure 4.10). This analysis identifies which neurophysiological traits remain stable across the recording duration and informs the patient-level aggregation strategy. The development-set median ICC of 0.435 (reported as a primary finding in Section 5.3.2) indicates moderate overall stability: this intermediate value confirmed that purely static aggregation (window means alone) would discard meaningful intra-recording variation. Trajectory-aware statistics—slope and IQR across the recording timeline—were therefore included alongside window-level means in the aggregation

layer. Aggregated features were persisted in columnar `Parquet` format to support $O(1)$ random-access retrieval during cross-validation.

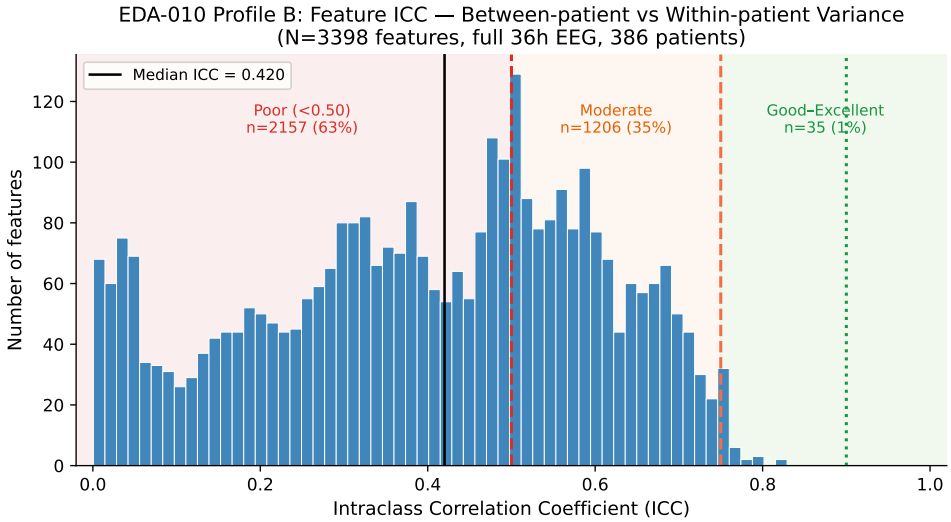



**Figure 4.10:** ICC distribution across the Profile B extended feature set ($N = 3{,}423$ features, comprising 19 monopolar and 18 bipolar channel derivations). Median ICC = 0.435, indicating moderate overall stability across windows. The 1,578 per-window features used in modelling are the monopolar subset of this set.

Extended ICC analysis, including domain-stratified categorisation and outcome-stratified variability, is documented in Section A.7.

### 4.4.3 Feature Computation Feasibility

Not all neurophysiological features were retained. Computationally expensive pairwise descriptors such as Sample Entropy and Coherence were disabled in the final thesis pipeline because they scale poorly across the full 32,000-hour corpus and would otherwise have dominated runtime. This pruning reflects a feasibility constraint rather than a claim that the descriptors lack value.

**Specparam failures as NMAR signal.** The handling of spectral parameterisation failures follows the NMAR protocol defined in Section 3.7. During feature extraction, windows on which `specparam` did not converge were not imputed. Instead, a binary `specparam_failed` flag was propagated as an explicit feature column. This flag is systematically elevated in highly suppressed and isoelectric segments, and propagated as an input feature to all three model families.

**Cardiac confound validation (EECI).** The feature set was planned to undergo a cross-check against the *EEG-ECG Contamination Index (EECI)* (Section 3.3.8) to rule out cardiac volume-conduction confounding before features entered Phase 1 modelling. Full-cohort EECI analysis was not completed within the scope of this thesis: the metric was implemented as a prototype, but its calibration required annotated reference windows that were not available in time. The unresolved question of cardiac contamination in the feature space consequently motivates the Profile C-noecg ablation in Phase 2 (Section 4.6) and is documented as future work in Section B.2.4.

## 4.5 Phase 1 Experiment: Feature-Based Machine Learning

Phase 1 implemented the feature-based baseline defined in Section 3.5. Sex classification and age regression were run on patient-level aggregates to evaluate the sex classification control and the primary CN-BAG feasibility hypotheses under strict subject isolation.

To prevent memorisation of patient-specific fingerprints, models were trained on patient-level aggregates rather than individual windows. The $\approx 10{,}994$-dimensional feature space (post-aggregation) was managed by enforcing *Complexity Bounds*: algorithms with $\mathcal{O}(N^2)$ complexity—Sample Entropy and full cross-channel coherence—were disabled after profiling showed they consumed $> 99\%$ of execution time. The retained algorithms were:

- *XGBoost (Gradient Boosted Trees):* Handles sparse high-dimensional inputs natively without imputation, allowing the `specparam_failed` convergence flag to enter as a first-class pathological indicator rather than missing data. Fixed hyperparameters are listed in Table B.15.
- *ElasticNet Regularisation:* A strictly linear baseline whose L1/L2 penalties enforce sparsity, mathematically pruning unstable temporal-drift covariates.
- *Random Forest (Ensemble Averaging):* An ensemble of decorrelated decision trees using Gini-impurity splits. No feature scaling is required; the tree structure is invariant to monotonic transformations. RF Gini-based feature importances serve as a cross-method consistency check against XGBoost gain scores.

Post-hoc interpretability was calculated via *TreeExplainer* (SHAP), using the C++ backend of the tree structures to efficiently compute marginal feature contributions. Model evalu-

ation occurred inside a five-fold `StratifiedGroupKFold` cross-validation scheme, stratified on age bins at the patient level and grouped by patient identity, ensuring zero exposure to the final 121-patient locked test set.

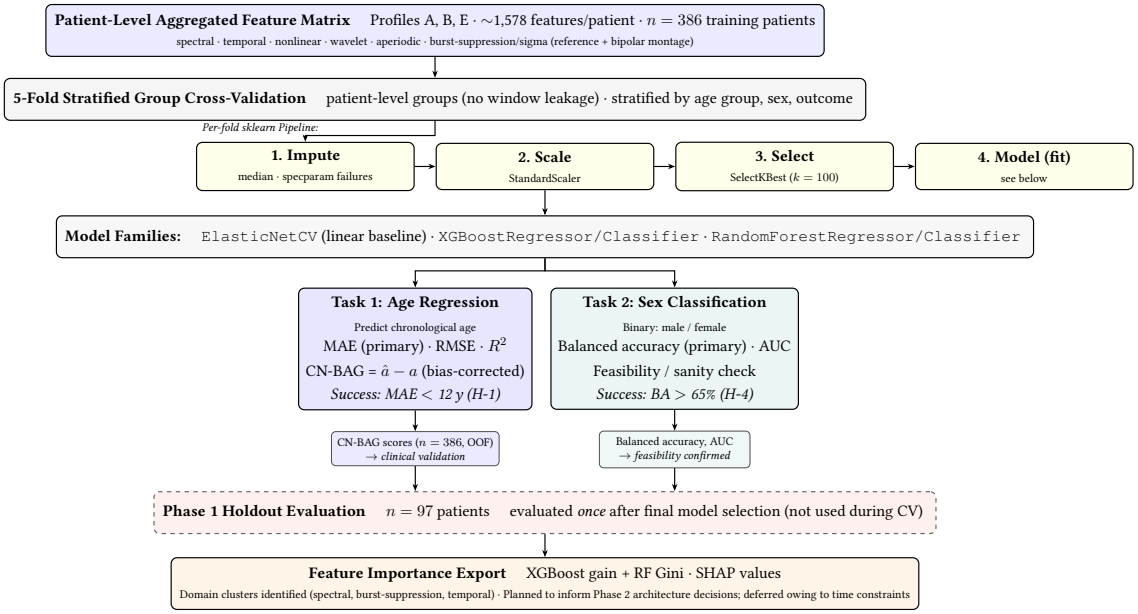



**Figure 4.11:** Phase 1 feature-based experiment pipeline. Patient-level aggregated feature matrices (Profiles A, B, and E) are fed into a five-fold `StratifiedGroupKFold` cross-validation loop. Each fold runs a full `sklearn` pipeline (imputation → scaling → feature selection → model) on two parallel tasks: age regression and sex classification. Three model families were evaluated: ElasticNet regularisation (linear baseline), XGBoost (gradient boosted trees), and Random Forest (ensemble averaging; Gini-impurity splits). The corrected CN-BAG output is derived from the age-regression head. SHAP and RF Gini importances were computed post-hoc and planned to inform Phase 2 architecture decisions; this cross-phase step was deferred owing to time constraints. The locked 121-patient test set is held out throughout.

**Age-bias correction.** Following the protocol defined in Section 3.5, post-hoc linear bias correction was applied to the raw CN-BAG outputs of all Phase 1 models. For each cross-validation fold, intercept $\alpha$ and slope $\beta$ were estimated from the training partition and subtracted as in Equation (3.3). Only the bias-corrected CN-BAG is used for downstream hypothesis testing.

## 4.6 Phase 2 Experiment: Time-Frequency Signal Deep Learning

Phase 2 implemented the end-to-end sequence-modelling track defined in Section 3.6. Phase 1 used Profiles A, B, and E feature matrices; Phase 2 ingested Profile C tensors to retain the continuous voltage stream for the sequence model. The waveform path operated on 10-second windows with 5-second overlapping steps (2,560 samples at 256 Hz), sized to fit within the available H100 GPU memory allocation. A matched ablation compared tensors with and without ECG removal:

- *Profile C (baseline tensor path):* Retains ECG volume conduction and applies only minimal filtering plus common-average referencing.
- *Profile C-noecg (matched ablation):* Adds targeted ECG-ICA and OBS fallback removal to the same baseline path, providing the cardiac-removal comparison defined in Section 3.3.4.

This comparison isolates whether ICA-based cardiac removal materially changes downstream representation learning.

Phase 2 did not jump directly to a single canonical network. An architecture pilot first compared four candidate architectures under the same pilot regime: `CNNBaseline`, `MultiTaskBrainAge`, `TCNMultiTask`, and `EEGConformerMultiTask`. Both the pilot and the subsequent canonical run (EXP-201) operated on the 286 development patients whose Profile C recordings met the Stage 1 quality threshold (out of 483), with a window budget of 50 windows per patient—a constraint imposed by GPU memory and SLURM wall-time limits. The goal was to test whether stronger temporal inductive biases — recurrent context with attention, dilated temporal convolutions, or transformer-style self-attention — materially altered the ceiling relative to a shared CNN encoder. The pilot comparison showed only small margins between architectures (mean validation age MAE 12.31–13.18 y; Table B.16), while the auxiliary sex task remained at chance for all models (BAcc 0.500–0.505). `MultiTaskBrainAge` was therefore carried forward to the canonical Phase 2 run (EXP-201) because it achieved the lowest mean validation age MAE while remaining the closest match to the temporal-transition hypothesis motivating Phase 2. A CPC 1–4 sensitivity run (EXP-201-excpc5) was additionally planned to probe whether restricting training to neurologically intact patients recovers a stronger age signal; the rationale for this cohort restriction, and its interaction with WLST-driven outcome labelling, is developed in Section 6.4.1.

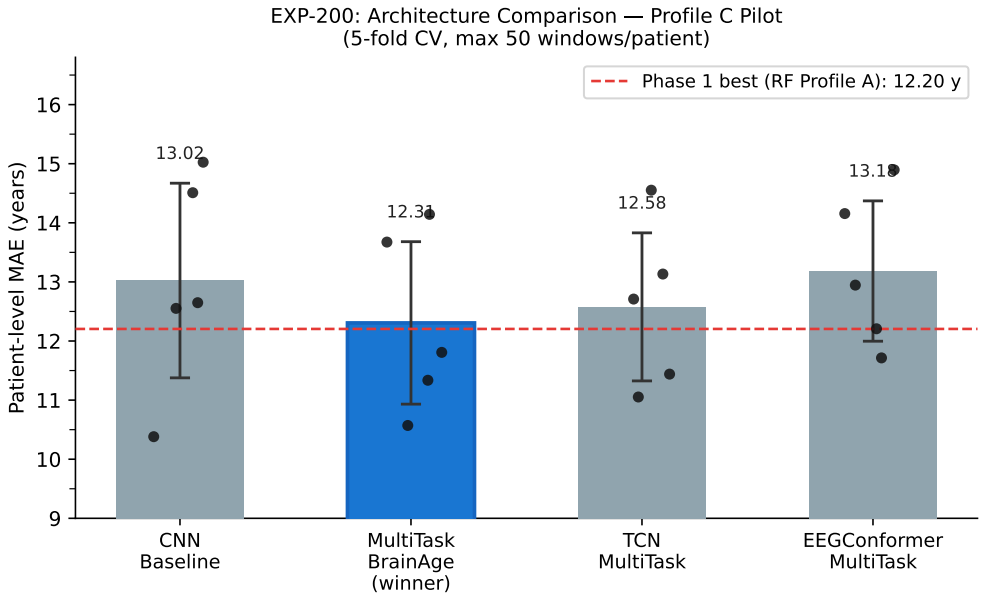



**Figure 4.12:** Architecture pilot screen (EXP-200). Four candidate architectures are compared under matched training conditions on Profile C tensors ($n = 286$ development patients, 50 windows per patient). The mean validation age MAE spans 12.31–13.18 y across architectures, and auxiliary sex balanced accuracy is at chance (0.500–0.505) for all, indicating that none recovers a meaningful sex trait in this setting. `MultiTaskBrainAge` (CNN-BiLSTM-Attention) was selected for the canonical run on the basis of its lowest mean validation age MAE.

The architecture (Figure 3.6) uses a CNN encoder as a temporal compressor before the BiLSTM and attention layers, reducing sequence length sufficiently to fit the available GPU memory budget while preserving a broad temporal receptive field. The multi-task design (simultaneous age regression and sex classification) forces the shared encoder to learn a representation that is not solely optimised for the age objective.

Model evaluation used a five-fold `StratifiedGroupKFold` cross-validation scheme, stratified on age bins at the patient level and grouped by patient identity to ensure strict subject isolation across folds. No locked test-set exposure occurred during this phase. Architecture and training hyperparameters for the canonical run are listed in Table B.8 (Section B.10).

Early stopping (patience = 10 epochs on validation MAE; Table B.8) serves as the primary regulariser. The patience is set conservatively to allow the cosine-annealing scheduler to complete at least one full restart before termination; the best checkpoint by validation MAE is retained rather than the final epoch.

### 4.6.1 GPU Determinism and Bit-wise Integrity

Ensuring bit-wise reproducibility across the tensor payload required addressing a subtle and frequently overlooked source of non-determinism in GPU computing. Standard *CUDA atomic-add operations* are non-deterministic by design: when multiple GPU threads accumulate gradients simultaneously, the specific order of floating-point additions varies between runs due to thread scheduling. Because floating-point arithmetic is not associative ($(a + b) + c \neq a + (b + c)$ in finite precision), microscopic round-off differences accumulate across thousands of iterations, causing identical model architectures to produce diverging results on the same data.

For this study, such non-determinism is scientifically unacceptable: a null result could otherwise be attributed to computational instability rather than a genuine biological limit. Strict GPU determinism was enforced via `CUBLAS_WORKSPACE_CONFIG=:4096:8` and `torch.use_deterministic_algorithms(True)`, at the cost of a modest runtime penalty. This ensures every result in Chapter 5 is perfectly reproducible regardless of the specific HPC node. Reproducibility is further guaranteed at the experiment level via the Git-hash injection described in Section 4.8.

## 4.7 Dataset and Infrastructure

The I-CARE cohort ($N = 604$; full characterisation and inclusion criteria in Section 3.7) forms the operational basis for all experiments. At ingestion, all recordings were standardised to the 19 standard 10-20 electrode positions common across all consortium acquisition sites; recordings that include additional non-EEG channels (ECG, $SpO_2$, utility) contributed only the 19-channel EEG subset to preprocessing and feature extraction. Patient flow through the development–test split and the Phase 1 sub-split is illustrated in Figure 3.3.

**Compute infrastructure.** All experiments were executed on the GWDG KISSKI HPC allocation (Section C.2.1). Preprocessing and Phase 1 modelling utilised CPU compute nodes; Phase 2 deep-learning training ran on dedicated GPU nodes (A100 / H100). Aggregate SLURM-allocated resource estimates across all preprocessing, feature-extraction, and modelling runs totalled approximately 166,000 CPU core-hours and 9,700 GPU-allocated core-hours (four concurrent GPUs per Phase 2 array job); actual CPU utilisation was approximately 10% owing to the sequential per-patient processing structure (Section C.2.1).

## 4.8 Continuous QA and Validation

Quality assurance was implemented as a multi-tier protocol covering data integrity, pipeline correctness, and scientific validity. Continuous monitoring was surfaced through interactive dashboards rather than deferred to post-run inspection. The quality gates targeted two distinct concerns: verifying the input data (metadata completeness, signal integrity) and verifying that the pipeline was operating correctly at each stage (notch filter frequency response, mask bit coverage, feature distributions).

**Data integrity checks.** The I-CARE archive was audited for metadata completeness (age, sex, CPC labels), channel counts, and raw signal integrity prior to any processing. The 3 patients with missing metadata labels (of 607) were excluded at this stage, yielding the working cohort of $N = 604$ patients (Table 4.1). No further patient-level exclusions were performed: all 604 patients contributed preprocessed data to the pipeline.

**Pipeline correctness validation.** Each preprocessing step was validated before full-cohort execution:

- The prominence-gated notch filter frequency response was plotted and visually confirmed to achieve >40 dB attenuation at the target notch frequencies without introducing roll-off artefacts in the adjacent neural bands.
- Quality mask bit coverage was validated on a 20-patient development probe to confirm that the bitmask thresholds captured genuine hardware faults without over-flagging pathological bursts.
- Preprocessing outputs were inspected per-profile at each step via a bespoke **interactive pipeline evaluation dashboard** (Panel/Bokeh; `04-code/05-visualization/pipeline_evaluation_dashboard.py`) that displayed per-patient signal traces, quality-mask bit coverage, PSD, per-segment quality metrics, and ECG-correlation summaries. This enabled stepwise visual confirmation of each profile's processing chain before parameters were frozen for full-cohort execution.
- Artefact characterisation for the quality mask was validated through a dedicated suite of executed analysis notebooks (`EXP-305c`), covering flatline detection, saturation, signal jumps, DC drift, and overall mask-effect summaries.
- Feature distributions were profiled after aggregation to detect NaN propagation, range violations, and unexpected collapses in inter-patient variance.

**Sex classification diagnostic.** As specified in Section 3.2.2, biological sex classification was retained as a diagnostic check rather than a hard go/no-go gate. The near-chance performance observed in the comatose cohort (reported in Chapter 5) was used to contextualise the age-regression findings, but it did not stop model training.

**Experiment-level reproducibility.** To guarantee that every result in Chapter 5 is exactly reproducible, the Git commit hash of the codebase was automatically injected at experiment time into every output artefact. Each SLURM submission script resolved the active commit hash via `git rev-parse HEAD` and passed it as an environment variable to the pipeline entry-point, which propagated it to every Parquet metadata header and every MLflow run record. This ensures that any reported result is traceable to the exact line of code that generated it: given the same commit hash, input data, and random seed (globally fixed at 42), results can be reproduced bit-for-bit on any compatible compute node.

### 4.8.1 Testing Strategy

The codebase employs a three-tier testing strategy to catch errors at different granularities.

**Unit tests.** Critical modules (preprocessing, feature extraction, model training) maintain $\geq$80% line coverage. Test execution is automated via `pytest`; failed tests block experiment submission to the HPC cluster.

**Integration tests.** Before each HPC submission, the full pipeline is run end-to-end on a 10-patient subset across all five profiles. A second integration test on a 50-patient subset verifies the target cluster architecture. These catch inter-module interface errors that unit tests miss.

**Post-experiment validation.** After each HPC run, output files are sampled for: correct column presence and types, physiologically plausible value ranges (PSD $\geq 0$, burst-suppression ratio $\in [0, 1]$), absence of unexpected NaN values, and demographic consistency across profiles for the same patient.

## 4.9 HPC Orchestration and Project Workflows

All experiments were executed via SLURM batch scripts on the GWDG KISSKI HPC allocation. The project used a Directed Acyclic Graph (DAG) to manage module dependencies; the full experiment dependency graph is shown in Section B.4. Experimental stages (Phase 1/2) were only unsealed once the preceding preprocessing and feature extraction stages passed their respective quality gates. To minimise I/O overhead across the raw data throughput, a *Load-Once/Fork* strategy was applied: raw waveforms were loaded into shared memory exactly once and then parallelised across profile-specific branches.

### 4.9.1 Continuous Monitoring and Interactive Dashboards

Analytical runs were tracked in a central **MLflow** instance, associating every recorded result with its Git commit hash and configuration matrix version to guarantee reproducibility (Section 4.8). Two bespoke interfaces supported continuous progress monitoring throughout the project (Figure 4.13 and Section C.4.2). These supported day-to-day

pipeline inspection and task coordination without requiring a backend service, and made a large, decentralised workflow inspectable while it was still being built and debugged:

- **Task Dashboard** (Figure 4.13): A serverless web application that parsed decentralised JSON files (`00-TASKS.json`) from each module directory, rendering a live Kanban board with status, priority, and dependency metadata. This supported day-to-day coordination of debugging, implementation, and writing work without requiring a backend service.
- **Module Dashboard** (Section C.4.2): Fed directly by the `EXPERIMENTS.md` single source of truth, this dashboard exposed a data-flow matrix showing which artefacts each pipeline module produced or consumed, the integration status of every experiment, and control-plane operation summaries. In practice, it functioned as a continuously updated architectural map of the whole project.

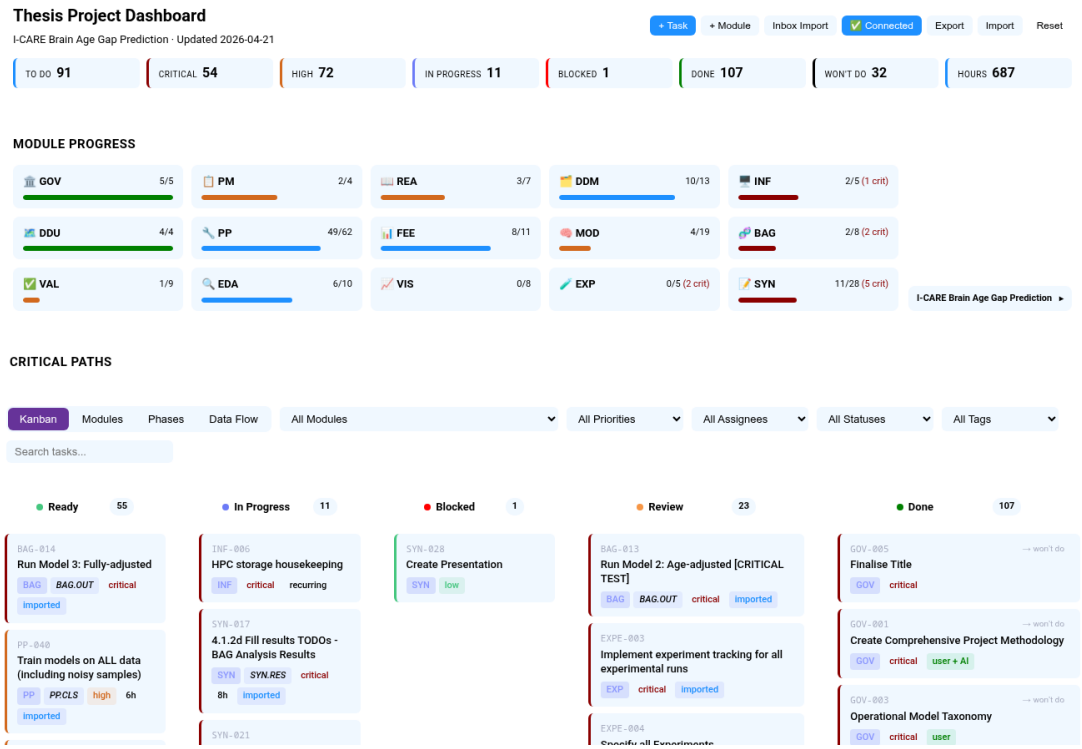



**Figure 4.13:** The `task_dashboard.html` Kanban board showing task status across all modules. Cards carry status, priority, submodule tags, estimated hours, and dependency flags. The aggregate summary row reports totals at final thesis submission.

# Chapter 5

# Results

Building on Chapter 4, this chapter reports the findings of the study. It begins with the main test-set and cross-validation results, then moves through preprocessing, feature engineering, and the two modelling tracks. Extended tables, fold summaries, and the Hypothesis Outcome Matrix are provided in the appendices (Sections B.1, B.14, B.14.1 and B.16).

## 5.1 Primary Findings

Across all modelling tracks, the study produced the same broad pattern. The five primary results are:

1. **Primary Test for CN-BAG (EXP-202):** On the overall locked test set, the final deep learning evaluation yielded a best Mean Absolute Error (MAE) of 12.90 years ($R^2 \approx 0$), failing to track chronological age.

2. **Biological sex diagnostic:** Binary sex estimation remained near chance in both phases. Phase 1 cross-validation yielded a best balanced accuracy (BAcc) of $0.577 \pm 0.077$ (ElasticNet, Profile A); the Phase 2 sealed test (EXP-202) returned BAcc $= 0.500$, predicting male for every patient.

3. **Sensitivity Test (nocpc5):** Restricting the evaluation to the survivor cohort (CPC 1–4, $N = 162$) yielded an apparently reduced Phase 1 MAE of 10.20 years (best model: XGBoost, Profile A). However, this subgroup has a younger age distribution (mean 58.1 y, median 58 y, $\sigma = 13.7$ y, IQR 50–68 y) compared to the full Phase 1 training cohort (mean 60.8 y, median 62 y, $\sigma = 15.6$ y, IQR 51–72 y), and the

corresponding null-model threshold is also lower (10.92 y vs. 12.61 y). The apparent error reduction therefore reflects this arithmetic distribution shift rather than true biological trait recovery. Sex classification similarly failed at chance level in this subgroup (best CV BAcc $= 0.548$, AUC $= 0.598$).

4. **Identity verification (holdout vs. test):** Intra-patient temporal holdout evaluation demonstrated near-perfect memorisation (Sex BAcc 0.988, Age MAE 10.13 years), while inter-patient generalisation collapsed. The holdout/test contrast shows that the models captured patient-specific structure without yielding generalisable age or sex prediction.

5. **Intraclass Correlation (ICC) drift:** Feature stability analysis revealed a median ICC of 0.435 over the monitoring window, showing substantial within-patient temporal variability relative to between-patient stability.

These findings are unpacked in the remaining sections. The chapter moves from signal recovery and feature behaviour to the Phase 1 and Phase 2 modelling results, before turning to CN-BAG outcome analyses. Chapter 6 then focuses on interpretation. The complete Hypothesis Scorecard summarising the formal verdicts for H-1 to H-11 is provided in Table B.1 (Section B.1).

### 5.1.1 Age Prediction and Diagnostic Outcomes

The failure to extract age and sex is visualised in the test set outcome distributions. Figure 5.1 demonstrates the constant-predictor collapse where the model defaults to the population mean, while Figure 5.2 highlights the chance-level performance of the sex classifier.

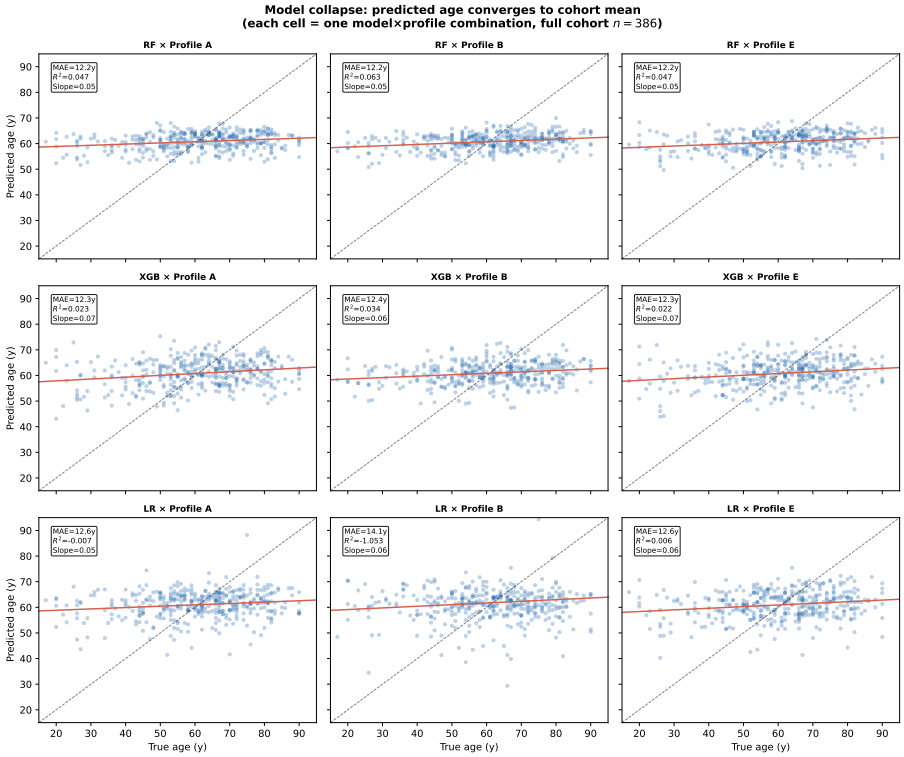



**Figure 5.1:** The predictive collapse of the models. The scatter distribution shows an $R^2 \approx 0$, where predictions default to the cohort mean ($\approx$ 61 years) regardless of the patient's actual chronological age.

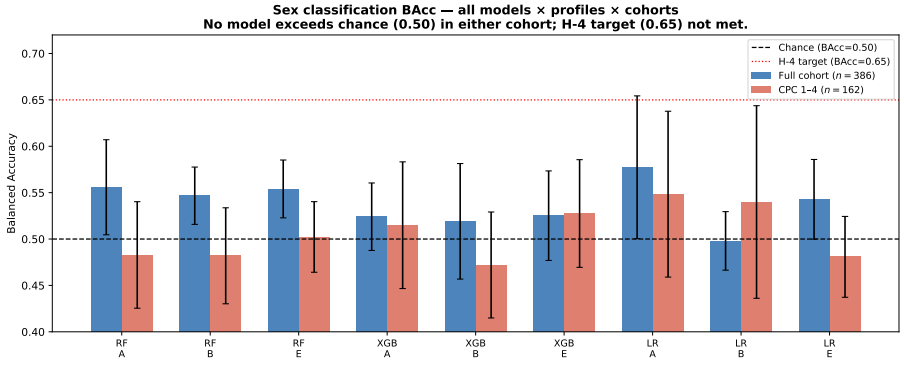



**Figure 5.2:** Balanced Accuracy for biological sex estimation across models. The near-chance performance confirms that even fundamental binary traits are obscured by the comatose state.

## 5.2 Preprocessing Results: Signal Recovery

### 5.2.1 Temporal Window Exclusion Characteristics

Profile E exhibited a substantially lower hardware-fault exclusion rate than Profile A (median window-pass fraction: $0.968$ vs. $0.669$; Figure 4.8 and Section 4.3.1). As discussed in Section 4.3.1, The quality masking detects time-domain hardware failures only; Profile E's lower exclusion rate also partly reflects that upstream step-removal reduces the amplitude transients triggering those criteria. Data passage does not certify freedom from physiological artefacts. The spatial and temporal pattern of bitfield activations on a representative recording is shown in Figure C.7b.

### 5.2.2 Failure of Normative Algorithms

As detailed in Section 4.3.1, the industry-standard PREP/RANSAC bad-channel detector proved fundamentally incompatible with comatose EEG: global delta/theta synchrony violates the algorithm's core assumption that a defective channel will be poorly predicted from its neighbours, rendering the spatial criterion unreliable and necessitating hardware-physics thresholds instead.

Notwithstanding the lower exclusion rate among profiles, cross-profile age-prediction MAE remained statistically indistinguishable across all preprocessing variants ($|\Delta\text{MAE}| \leq 0.01$ y, Wilcoxon $p_{\text{adj}} = 1.000$). The complete cross-profile comparison for H-6 is given in Section B.17.6 (Table B.27).

## 5.3 Feature Representation Quality

### 5.3.1 Bimodality and the Specparam Failure

A detailed analysis of the 1,578-dimensional feature space revealed a $42.6\%$ prevalence of bimodal feature distributions. This bimodality is a direct neurophysiological consequence of burst-suppression pathology. Furthermore, the standard spectral parametrisation algorithm (`specparam`) failed to fit a valid $1/f$ slope on $49.2\%$ of segments during deep suppression. Crucially, this failure was *Not Missing At Random (NMAR)*; the missingness itself became a high-importance split in the tree-based models, correctly identifying severe injury states rather than age.

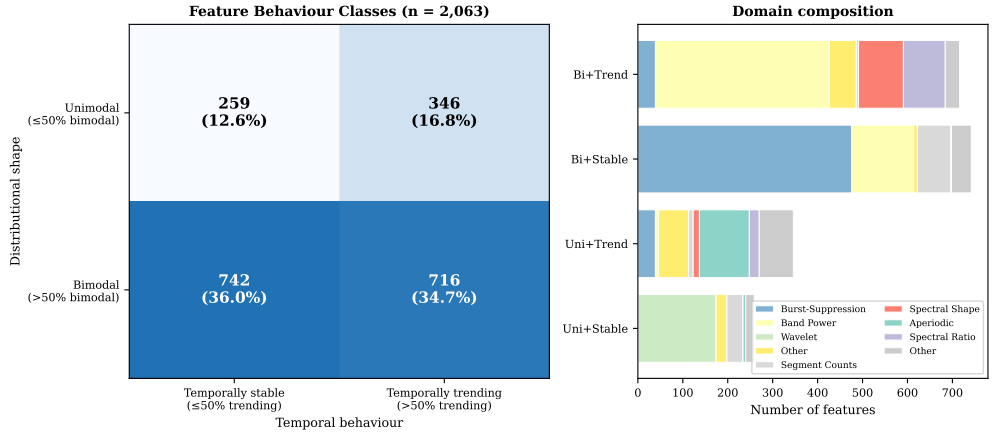



**Figure 5.3:** Cross-tabulation of bimodality and monotonic temporal-trend status across the 1,578-feature modelling matrix (Profile B). Features are classified along two axes: whether their distribution is bimodal (indicating burst-suppression–driven behaviour) and whether they exhibit a significant monotonic temporal trend across the 36-hour window. The joint structure confirms that bimodality and temporal drift are partially independent phenomena: a feature can be bimodally distributed but temporally stable, or unimodal but strongly drifting, motivating the use of separate aggregation strategies rather than a single global rule.

### 5.3.2 Temporal Drift and Intraclass Correlation (ICC)

As noted in the primary findings, the median ICC was $0.435$ (see Figure 4.10). The ICC summary refers to the extended 3,423-feature Profile B set; the 1,578-feature modelling matrix is its monopolar subset. Significant monotonic temporal trends were observed in 89.0% of features in more than 30% of patients, and in 68.4% of features in a strict majority (>50%) of patients across the 36-hour measurement window (see Figure 4.9). The implication is clear: within-patient variance (evolving injury) dominates the static trait signal.

The broader age/sex confound structure is shown in Figure A.14 in Appendix A. Age-related and sex-related correlations occupy only partially overlapping feature domains, reinforcing the decision to treat sex as a positive control rather than as a reliable auxiliary biomarker for stable trait recovery.

### 5.3.3 ECG Contamination Trade-off

Benchmarking of cardiac-artefact removal showed that any method achieving $> 30\%$ contamination reduction simultaneously destroyed $> 50\%$ of the prognostic neural delta power. The implications of this trade-off are taken up in Chapter 6.

## 5.4 Phase 1 Results: Feature-Based Machine Learning

Phase 1 tested whether handcrafted feature matrices yielded cross-patient recovery of age or sex from the coma EEG. Table B.6 (Section B.10) provides an at-a-glance summary of all Phase 1 configurations evaluated, their feature profiles, and their cross-validation performance against the null baseline; the subsections below discuss the results in detail.

### 5.4.1 Age and Sex Baseline Collapse

Across XGBoost, Random Forest, and Elastic Net, the cross-patient age prediction defaulted to the cohort mean ($MAE \approx 12.2$y, $R^2 \approx 0$). Similarly, sex classification BAcc hovered between $0.50$ and $0.58$. Full model-grid results are reported in Section B.17.1 (Table B.22) for age regression and Section B.17.5 (Table B.25) for sex classification.

The corresponding CPC 1–4 sensitivity-cohort results are in Table B.23 and Table B.26 respectively.

The sensitivity analysis restricting the evaluation to CPC 1–4 patients ($N = 162$, survivors only) produced an apparently lower Phase 1 best MAE of 10.20 years (XGBoost, Profile A). This reduction is, however, an arithmetic consequence of the subgroup's younger age distribution (mean 58.1 y, $\sigma = 13.7$ y; null-model threshold 10.92 y) rather than any genuine improvement in biological trait recovery: when the population is younger and its variance smaller, the constant-predictor baseline is proportionally easier to approach. Figure 5.4 shows this cohort-composition shift directly. The CPC 1–4 subgroup is younger and narrower than the full Phase 1 development cohort, whereas the locked final test set broadly tracks the full development distribution. Figure 5.5 then shows an illustrative full-versus-survivor comparison for the best full-cohort Phase 1 configuration (Random Forest, Profile A): predictions remain tightly collapsed around the cohort centre in both cohorts, even where the subgroup MAE numerically improves.

### 5.4.2 Identity Verification: The Memorisation Gap

To ensure the models were not simply underfitting due to noise, an intra-patient temporal holdout diagnostic (Level 1 validation) was evaluated. On these holdout windows, age MAE dropped to $10.13$y (XGBoost) and sex BAcc surged to a near-perfect $0.988$ (Random Forest). This large *memorisation gap* between intra-patient holdout performance (Sex BAcc 0.988) and cross-patient generalisation (chance-level BAcc $\approx 0.58$) shows that the models captured patient-specific spatial-spectral structure across the monitoring window

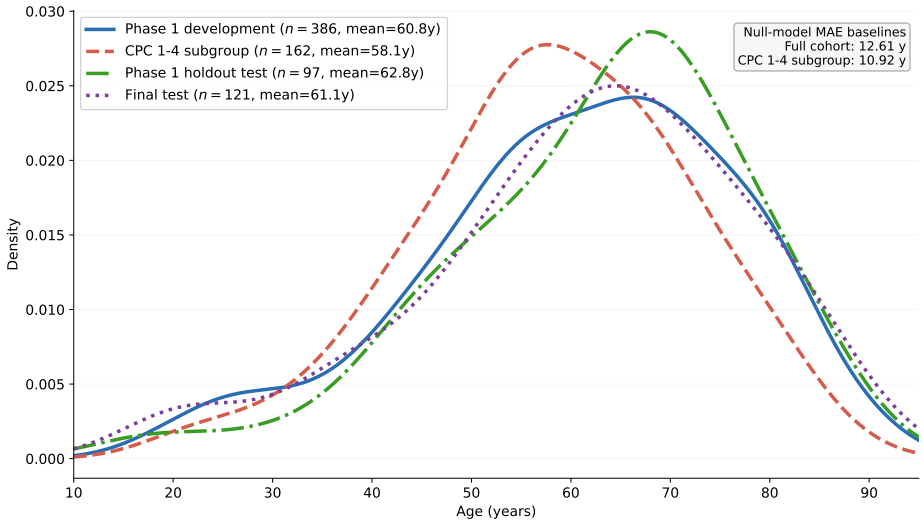



**Figure 5.4:** Kernel density estimates of age across the main analysis cohorts. The CPC 1–4 survivor subgroup is shifted younger and has a narrower spread than the full Phase 1 development cohort, whereas the Phase 1 holdout and the locked final test set broadly track the full development distribution. The panel annotation reports both null-model MAE baselines used in the sensitivity interpretation (full cohort: 12.61 y; CPC 1–4 subgroup: 10.92 y). Because the null-model MAE scales with cohort age variance, this narrower survivor-only distribution lowers the subgroup baseline and makes apparent sensitivity-cohort improvements easier to obtain even when the underlying predictor remains collapsed.

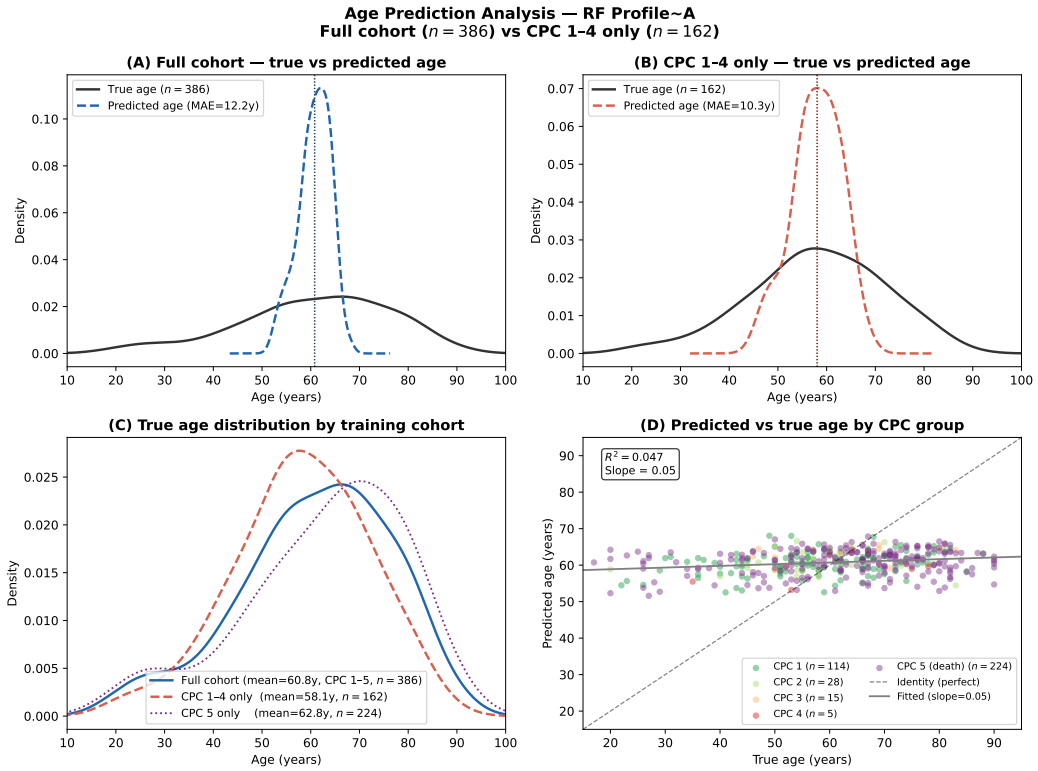



**Figure 5.5:** Illustrative sensitivity-cohort comparison for Random Forest, Profile A. Panels A and B show predicted-age collapse toward the cohort centre in both the full development cohort (MAE 12.2 y) and the CPC 1–4 survivor subset (MAE 10.3 y). Panel C visualises the same age-distribution shift by cohort, and Panel D shows the near-flat predicted-versus-true relation across CPC groups (fitted slope 0.05). The lower survivor-only MAE therefore reflects a narrower target distribution rather than genuine recovery of age-tracking behaviour.

but did not translate that structure into generalisable age or sex prediction for unseen patients. This memorisation gap is quantified in Figure 5.6.

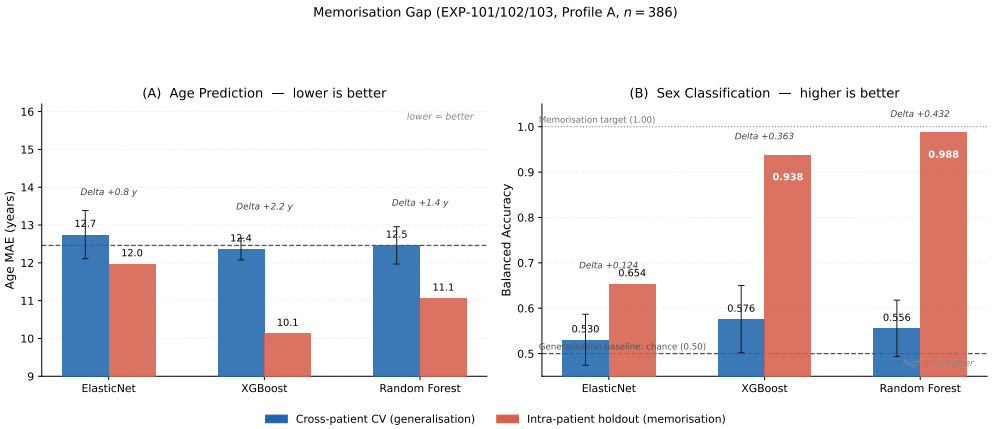



**Figure 5.6:** The Memorisation Gap across all three Phase 1 architectures (ElasticNet, XGBoost, Random Forest). *(A)* Age MAE and *(B)* balanced accuracy for biological sex classification, comparing cross-patient test-set generalisation (blue, with 95% CV intervals) against intra-patient temporal holdout (red, Level 1 validation). All three models memorised unique spatial-spectral biometric fingerprints (Sex BAcc 0.654–0.988 holdout) but collapsed to near-chance on unseen patients (BAcc 0.530–0.576), confirming state-trait inseparability rather than model underfitting. Null-model baseline (MAE = 12.46 y) shown as a dashed line in *(A)*; in *(B)*, the dashed line marks chance-level generalisation (0.50) and the dotted line marks the intra-patient memorisation target (1.00).

The quantitative overfitting gap for each model family is tabulated in Section B.18.2 (Table B.31). Section B.18.1 (Table B.30) lists the top-10 features driving the age regression, confirming the models anchored on acute-state markers (spectral edge frequency, delta power) rather than stable age-related traits.

SHAP-based feature-importance analysis (H-7 severity validation) was evaluated; however, with $R^2 \approx 0$, the model collapses to predicting the population mean and SHAP attributions reflect noise rather than systematic feature contributions. The Jaccard feature-overlap check between the age model and a CPC outcome model is therefore not interpretable for this null result.

## 5.5 Phase 2 Results: Time-Frequency Signal Deep Learning

Phase 2 tested whether end-to-end representation learning could bypass the feature bottleneck. All Phase 2 runs used Profile C waveform tensors (magnitude spectrograms,

0.5–30 Hz). Table B.7 (Section B.10) summarises all configurations evaluated; the subsections below discuss the results.

### 5.5.1 Architecture Comparison (EXP-200) and Final Evaluation (EXP-202/203)

Despite the increase in architectural complexity, Phase 2 deep learning models yielded statistically indistinguishable results from Phase 1. EXP-200 selected `MultiTaskBrainAge` as the strongest candidate, but the advantage over the other waveform models remained modest: mean validation age MAE ranged from 12.31 y (`MultiTaskBrainAge`) to 13.18 y (`EEGConformerMultiTask`), while all auxiliary sex scores stayed at or near chance (Table B.16). This cross-phase near-equivalence is summarised in Figure 5.8: even the strongest Phase 2 pilot and full runs remain clustered around the best Phase 1 errors and stay above the H-1 target of MAE $< 12$ y. On the final locked test set (EXP-202), the model achieved an MAE of $12.90$ y (95% CI $[10.73, 15.24]$) and a sex classification BAcc of $0.500$ — the model predicted male for every patient, the majority class in the 68.9% male training cohort. This mirrors the Phase 1 sex diagnostic on the waveform-based path. Full Phase 2 fold summaries (including the per-fold sex BAcc of $0.500 \pm 0.000$), the architecture comparison grid, and the formal H-11 verdict are provided in Section B.14.1 (Tables B.16 and B.17) and Section B.17.11.

Figure 5.7 shows the training dynamics underlying these summary results. All four architectures converge within 30 epochs: training age-MSE drops from approximately 260–280 at epoch 1 to approximately 25–35 at epoch 30, confirming that the models are not collapsing to a trivial constant predictor during training. The validation loss tracks the training loss throughout, with no prominent divergence indicative of overfitting. The patient-level MAE figures in the legend represent the post-training fold-evaluation scores (the authoritative CV metric), which are higher than the within-epoch monitoring values owing to the difference between window-level training targets and the patient-median aggregation applied at evaluation time.

**CPC 1–4 sensitivity run (EXP-201-excpc5).** A CPC 1–4 sensitivity run (EXP-201-excpc5) tested whether restricting the development cohort to neurologically intact and moderately injured patients reduced age estimation error. Of the 286 quality-filtered development patients, those with CPC 5 outcomes were removed, yielding $N = 129$

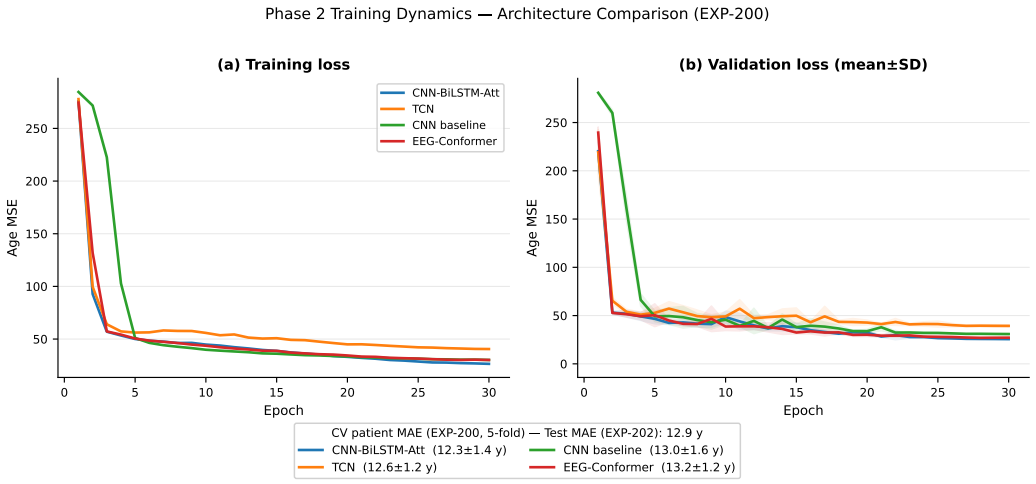



**Figure 5.7:** Phase 2 training dynamics for the four candidate architectures evaluated in EXP-200 (5-fold cross-validation on the 286 quality-filtered development patients drawn from the 483-patient development set; mean ± SD across folds). *(a)* Training age-MSE loss. *(b)* Validation age-MSE loss; the legend reports the corresponding post-training patient-level CV MAE (the authoritative metric; see also Table B.16). All models converge within 30 epochs without prominent overfitting. The final locked test-set evaluation (EXP-202, $N = 73$ patients) yielded MAE $= 12.90$ y, consistent with the cross-validation estimates of 12.3–13.2 y.

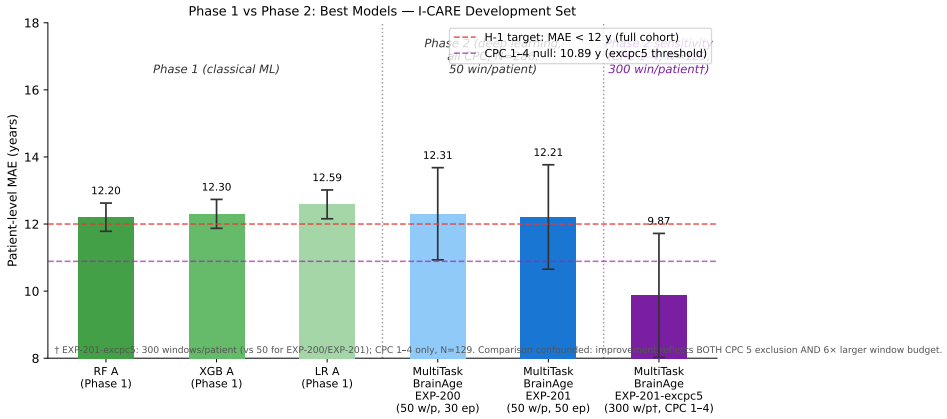



**Figure 5.8:** Best-performing development-set models from Phase 1 and Phase 2. Bars show patient-level MAE; error bars denote between-fold standard deviation. Phase 1 and Phase 2 full-cohort configurations (EXP-200 and EXP-201; 50 windows per patient) all remain above the H-1 target of MAE $< 12$ y. The rightmost bar (EXP-201-excpc5, purple) shows the CPC 1–4 sensitivity run ($N = 129$ quality-filtered patients), which achieved MAE $= 9.87$ y, below the CPC 1–4 subgroup null of 10.89 y (purple dashed line). However, this run used 300 windows per patient (vs 50 for all other Phase 2 bars), so the improvement conflates CPC 5 exclusion with a sixfold increase in per-patient training data; the comparison is therefore confounded and not directly interpretable as evidence of outcome-group effects.

patients whose Profile C recordings met the hardware-fault threshold. Unlike EXP-200 and EXP-201, which applied a per-patient window cap of 50, this sensitivity run used a cap of 300 windows per patient (the HPC default configuration), providing approximately six times more training signal per patient. The cross-validated patient MAE was $9.87 \pm 1.85$ y (fold range: 7.97–12.66 y), below the CPC 1–4 subgroup null of 10.89 y (mean absolute deviation from group mean; development CPC 1–4 subgroup, $N = 203$) at the point estimate. Sex classification remained at chance (BAcc $= 0.500$ across all folds). The apparent improvement relative to the full-cohort EXP-201 result ($\Delta = 2.34$ y) does not reach conventional significance at five folds (Wilcoxon signed-rank $p = 0.156$, Cohen's $d = 0.69$), reflecting the low statistical power of fold-level paired comparisons. Critically, this run conflates two simultaneous changes: removal of CPC 5 patients (whose EEG age estimates are expected to be most attenuated) and a sixfold increase in per-patient training windows. These factors cannot be disentangled in the present design, so the result is best treated as a directional exploratory signal rather than a controlled test of outcome-group effects on brain age estimation. To isolate the CPC 5 exclusion effect cleanly, a matched run with 50 windows per patient on the CPC 1–4 cohort would be required. This result is treated as a sensitivity signal rather than a hypothesis-confirming reduction; it is discussed further in Section 6.4.1.

## 5.6 Outcome Association and Sensitivity Analyses

### 5.6.1 CN-BAG Outcome Association

The CN-BAG analysis uses out-of-fold (OOF) predictions from the Phase 1 model trained on all 386 development patients (CPC 1–5 combined); each patient's Brain Age Gap therefore derives from a fold in which they were held out, preventing leakage. In the primary analysis, the bias-corrected CN-BAG showed a non-significant trend toward association with neurological outcome (good: CPC 1–2, $n = 142$; poor: CPC 3–5, $n = 244$; Mann–Whitney $p = 0.058$). A sensitivity analysis excluding CPC 5 patients—in whom WLST confounding is most acute, since recorded death may follow a clinical care-withdrawal decision rather than biological injury alone—compared good (CPC 1–2, $n = 142$) versus poor neurological outcome among survivors (CPC 3–4, $n = 20$) and yielded a nominally significant bias-corrected CN-BAG difference ($p = 0.009$, AUROC $= 0.680$ [0.555, 0.801]). This result is hypothesis-generating rather than confirmatory: excluding CPC 5 reduces but does not eliminate WLST confounding, and the poor-

outcome group is severely underpowered ($n_{\text{poor}} = 20$). The raw uncorrected BAG showed a stronger apparent separation (CPC 1–2 vs. CPC 3–5: $p = 0.0019$), attributable to chronological-age imbalance between outcome groups rather than neurophysiological deviation (Section 6.4.1). Figure 5.9 shows the CN-BAG distributions using grouped CPC strata (CPC 1–2, CPC 3–4, CPC 5) for visual legibility.

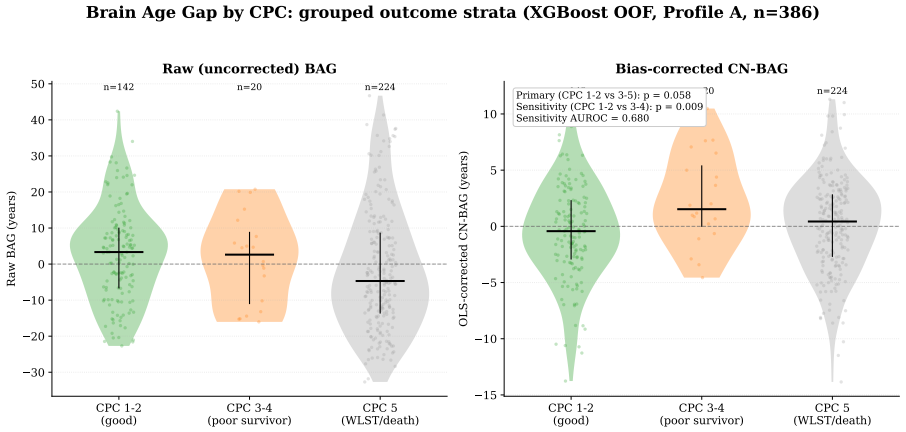



**Figure 5.9:** Brain Age Gap (BAG) distributions across grouped CPC outcome strata. Left panel: raw (uncorrected) BAG; right panel: OLS bias-corrected CN-BAG. In the corrected panel, the poor-outcome survivor group (CPC 3–4 combined, $n = 20$; median $+1.53$ y) has a higher corrected CN-BAG than the good-outcome group (CPC 1–2, $n = 142$; median $-0.43$ y), consistent with the sensitivity-analysis association ($p = 0.009$, hypothesis-generating). CPC 5 patients ($n = 224$) are shown separately; their inclusion in the primary analysis attenuates the association ($p = 0.058$) owing to WLST confounding.

Full CN-BAG outcome association statistics — Mann–Whitney U-test, Spearman $\rho$, and logistic regression coefficients — are provided in Section B.17.2 (Table B.24). The ordinal CPC gradient (H-3) and severity contrast (H-4) are in Section B.17.3 and Section B.17.4 respectively.

Chapter 6 interprets these findings, examining the biological and methodological factors that produce the predictive collapse, and positioning the result within the broader brain-age literature.

# Chapter 6

# Discussion

The results presented in Chapter 5 require interpretation across the operational dimensions that governed this thesis. This chapter treats the null result as the central finding, separates engineering success from biological limitation, and advances the state-trait interpretation as its central scientific claim. It follows the same broad order as the Results chapter: first the null result itself, then the preprocessing implications, then CN-BAG interpretation and its limits. Sections 6.1–6.4.1 develop the mechanistic interpretation of state-trait inseparability in this cohort; Section 6.5 locates the result within the brain-age literature; Section 6.6 catalogues the alternative explanations; and Sections 6.7–6.8 address methodological limitations and future directions.

Alongside the null result, three transferable contributions emerge from this work. First, the hardware-anchored quality-mask pipeline provides a transferable *quantify-not-remove* preprocessing strategy for high-pathology EEG. Second, the state-trait inseparability finding provides a falsifiable scientific boundary condition for normative modelling in acute coma rather than an unqualified performance claim. Third, the WLST-aware dual-analysis protocol and formal bias framing establish a reusable methodological and ethical guardrail for outcome-association studies in decisional-confounded ICU cohorts. These three strands also define a clear dissemination structure: a methodological paper on pathology-preserving preprocessing, a scientific paper on state-trait inseparability under acute HIE, and a methodological-ethical paper on WLST-aware bias control in neuroprognostic modelling.

## 6.1 Primary Findings: State-Trait Inseparability

The central finding of this thesis is that state-trait inseparability is the most plausible interpretation of the null results reported in Chapter 5 (cf. Section 3.1.1). Across all five primary findings, the acute phase of post-cardiac-arrest coma dominates the underlying biometric identity, rendering the underlying age trait unrecoverable with the current design. This interpretation remains falsifiable: a critical test is whether age/sex separability re-emerges longitudinally within the same patients as sedation is withdrawn and recovery trajectories unfold—a test that would require per-patient sedation-event timestamps not available in I-CARE.

### 6.1.1 The Substrate Recalibration

This study identifies a critical distinction from previous literature. Brain-age research in lighter sedation states, such as the elective surgical cohort evaluated by Sabbagh et al. [68], addresses a substantially lower-pathology signal regime than post-arrest coma. In contrast, hypoxic-ischaemic encephalopathy (HIE) appears to impose a substrate-level change: permanent neuronal injury and the physiological stress of targeted temperature management (TTM) plausibly shift the EEG so far away from the healthy or lightly sedated regime that the micro-variance needed for age discrimination is no longer available.

### 6.1.2 The Schmidt Cardiac Confound and Aperiodic Slopes

The failure of the aperiodic ($1/f$) features to recover age signal is not fully consistent with the recent finding by Schmidt et al. [9], which suggests that part of the age-related spectral slope in healthy adults may be a cardiovascular leak from the heart rate. In the ICU, where the cortical signal is suppressed by up to 90%, that confound is likely amplified. The practical reading is conservative: once the heart-rate component is controlled via aggressive ICA (Section 3.3.3) or overwhelmed by coma physiology, the remaining neural age signature appears too weak to support the regressor. This is corroborated empirically by the ECG trade-off benchmarking (Section 5.3.3): any removal method achieving $> 30\%$ cardiac contamination reduction simultaneously destroyed $> 50\%$ of the prognostic neural delta power, directly quantifying why aggressive cardiac unmixing is counterproductive in this signal. It should be noted that no evaluation was conducted with the full sealed test set for the ICA-cleaned Profile C-noecg; the comparison rests on cross-validation performance. Cardiac effects in brain data may also

be modulated by sedation and TTM, and the prominence-gated notch filter may have incidentally attenuated some residual cardiac contamination.

### 6.1.3 Outcome-Stratified ICC Difference

A small but consistent ICC difference was observed between outcome groups ($0.379$ vs. $0.455$). High temporal stability may indicate a more rigid or suppressed signal regime, whereas lower ICC may indicate preserved biological variability. This is suggestive rather than definitive, but it supports the broader feature-drift argument. The interpretation is reinforced by the temporal drift analysis: $89.0\%$ of features exhibited significant monotonic trends across the 36-hour ICC measurement window (Section 5.3.2), confirming that within-patient acute-state variance dominates the static trait signal.

The broader ICC landscape also has methodological implications for the aggregation design. Across the full Phase 1 feature set ($N = 3{,}423$ features, EDA-010-B, Profile B), $60.3\%$ of features fall in the poor ICC range (ICC $< 0.4$), $38.4\%$ moderate ($0.4$–$0.75$), and only $0.8\%$ achieve good ICC ($> 0.75$) (Figure A.8 in Appendix A). This breadth of poor within-patient consistency directly motivated the trajectory-aware aggregation strategy: static means would have erased meaningful distributional information for the majority of features.

### 6.1.4 Sex Classification as a Diagnostic Baseline

The ability to classify sex from the healthy EEG serves as a useful methodological baseline for critical care research. If an engineered machine learning pipeline fails to classify biological sex above random chance when applied to post-cardiac-arrest patients, it acts as a positive control for this feature set—suggesting that the pathological slowing and pharmacological suppression induced by severe HIE may be overwhelming the subtle, premorbid biological traits of the patient.

Phase 1 sex classification reached a best cross-validated BAcc $= 0.577 \pm 0.077$ (95 % CI [0.50, 0.63], AUC $= 0.600$, permutation $p = 0.001$; ElasticNet, Profile A)—statistically above chance but far below the 0.65 feasibility threshold. Phase 2 deep learning was fully degenerate (BAcc $= 0.500$ in all folds). The safest reading is that a weak residual demographic signal remains detectable in Phase 1, but not strongly enough to support a reliable trait measure in this cohort. Importantly, sex classification is used as a *diagnostic indicator*, not a hard gate: age regression training proceeds regardless of this outcome. The conservative conclusion is that the current representation is too compressed to

recover a reliable demographic trait, which a priori contextualises the age-regression null result.

The Phase 2 complete degeneration to BAcc $= 0.500$ reflects the 68.9% male cohort imbalance applied to an unweighted auxiliary loss, as detailed in Section 5.5: the minimum-loss strategy for an unweighted classifier on a 2:1 imbalanced dataset is to predict the majority class for every input.

### 6.1.5 Identity Verification and the Memorisation Gap

As reported in Section 5.4.2, intra-patient temporal holdout performance (Sex BAcc 0.988, Age MAE 10.13 years) vastly exceeded cross-patient generalisation (chance-level BAcc, MAE 12.9 years). This memorisation gap confirms that the pipeline captured high-fidelity individual biometric fingerprints; the failure to generalise across patients reflects state-trait inseparability rather than model underfitting or pipeline failure.

The verified train/test split integrity and the consistency of the null across both modelling tracks make it unlikely that the result is a simple artefact of computational failure.

### 6.1.6 Ruling out Technical Bottlenecks

As documented in Appendix C, the pipeline scaled across the available CPU and GPU infrastructure and reproduced the same null pattern across both modelling paradigms. The safest reading is therefore that the bottleneck in comatose brain-age estimation is neurophysiological rather than infrastructural.

The diagnostic feature rankings in Section B.18.1 provide a compact mechanistic summary of this pattern: the models prioritised acute-state markers such as spectral edge frequency and delta power rather than stable age-linked traits. This is reinforced by the 49.2% *specparam* failure rate (Section 6.7.2), a strong indicator of non-random missingness under severe cortical suppression. A related confound is that features such as burst-suppression ratio and aEEG amplitude correlate with both age and injury severity, meaning a model could in principle achieve a lower MAE not by recovering genuine age signal but by exploiting the covariance between injury severity and apparent EEG age — inflating any outcome association as a circular artefact rather than a biologically independent finding.

## 6.2 Signal Cleaning without Trait Recovery

The results reveal a simple but important pattern: successfully removing technical noise in the time domain does not, by itself, recover biological traits that have been suppressed or overwritten by the coma state.

Across all preprocessing intensity levels—from the unprocessed Profile A baseline to the fully processed Profile E—age-prediction MAE remained statistically indistinguishable ($|\Delta\text{MAE}| \leq 0.01$ y, Table B.27). This directly demonstrates that additional spectral processing does not unlock additional trait variance: the performance bottleneck is not technical noise in the time domain but the absence of recoverable age information in the frequency domain under acute coma. Critically, the cross-validation comparison of the ECG-removal ablation (Section 3.3.3) against the retained-ECG baseline (Section 3.3.3) showed no significant MAE improvement, suggesting that the variance removed during cardiac unmixing was genuinely non-neural and carried no recoverable age signal. This comparison is based on cross-validation results; no sealed test set evaluation was conducted for Profile C-noecg.

The hardware-grounded quality mask is nevertheless a directly transferable output of this thesis. Standard EEG preprocessing tools (PREP [40], ICLabel, MNE defaults) assume signal properties of healthy awake or sleep recordings. As demonstrated by the RANSAC failure (Section 4.3.1) and the empirical CDF analysis (Section 4.2.1), these assumptions are violated in post-cardiac-arrest coma, motivating the hardware-grounded alternative. The architecture successfully removes hardware artefacts while preserving pathological content; it cannot create age-related variance where none exists.

Profile A's comparable MAE also constrains one obvious alternative explanation. If the preprocessing merely removed variance that a regressor could exploit—for example, high-frequency noise that happened to correlate with age—then the cleaned profiles should have shown a clear MAE penalty. That Profile C does not do so suggests that the removed variance is genuinely non-neural, and that the residual signal is a more faithful characterisation of cortical activity. The failure to predict age therefore reflects a property of the underlying signal, not an artefact of overaggressive cleaning. Quantitatively, the Phase 1 test MAE is effectively invariant across profiles: Random Forest spans only 12.20–12.21 y across Profiles A, B, and E; XGBoost spans 12.30–12.44 y (Table B.22). The XGBoost training gap does narrow slightly under Profile B (from $+10.47$ y on Profile A to $+7.99$ y), suggesting that cleaner features reduce memorisation of patient-specific state patterns—yet this does not translate into any test improvement. The test floor is

therefore a property of the signal, not of the preprocessing configuration.

A converse risk warrants acknowledgement: some signals classified as artefact may carry clinical information. In post-cardiac-arrest patients, status myoclonus produces high-amplitude EMG contamination on scalp EEG that overlaps spectrally with technical muscle artefact. Myoclonus is a recognised poor-prognostic marker. The current quality mask annotates EMG-range activity as artefactual without distinguishing clinical myoclonus from non-pathological muscle contamination. Because the mask-based approach preserves the underlying signal rather than deleting it, the information is not lost; however, downstream feature extraction treats masked samples identically regardless of origin. Resolving this distinction would require expert neurophysiologist annotation—not available for I-CARE.

## 6.3 CN-BAG Interpretation

The OLS-corrected CN-BAG is a zero-centred deviation score: by construction of the residual correction, the population median is approximately zero ($+0.20$ years; IQR $-2.71$ to $+2.69$ y, $\sigma = 4.28$ y). A positive corrected CN-BAG indicates that a patient's BAG is higher than expected for their chronological age within this comatose cohort; a negative value indicates the reverse. The raw uncorrected BAG has a substantially wider distribution (median $-1.43$ y, IQR $-11.65$ to $+9.03$ y, $\sigma = 15.4$ y), reflecting the regression-to-mean artefact: because the model collapses to predicting near-constant age, the raw BAG distribution approximately mirrors the cohort age dispersion ($\sigma_{\text{age}} = 15.7$ y) and carries no additional neurophysiological signal.

What the corrected CN-BAG does *not* reveal is more important than what it does. An individual's deviation simultaneously encodes premorbid neurophysiology, the severity of the acute insult, pre-existing neurological conditions absent from the dataset, and sedation or TTM—all of which shift EEG profiles in directions that cannot be separated from EEG alone; this inseparability is a scientific finding, not a correctable confound (Section 6.4).

A responsible deployment would present CN-BAG as a population-relative summary statistic, comparable in form to a z-score, rather than a physiological measurement of any single underlying quantity. Whether a particular CN-BAG value crosses a clinically actionable threshold is an empirical question for properly powered prospective work, not the current retrospective cohort analysis.

With the metric's distributional properties and interpretive limits established, the

following section examines why those limits are not merely statistical but arise from the signal itself.

## 6.4 Confounder Inseparability

Recent longitudinal evidence suggests that the EEG-derived brain age gap is better characterised as a stable individual trait than as a marker of acute pathological state [6]: in subjects with transitions between pathological and non-pathological EEG, the brain age gap does not significantly change ($p > 0.4$), and direct EEG pathology classifiers consistently outperform brain-age-gap-based pathology detection. If the brain age gap is predominantly trait-like, then recovering it requires that premorbid inter-individual variation survives the acute perturbation.

In post-cardiac-arrest HIE, three sources of confounding operate simultaneously and produce overlapping EEG signatures: (i) *permanent cortical damage* from ischaemia and reperfusion injury, which irreversibly alters the neural substrate; (ii) *reversible pharmacological and physiological state effects*—sedation (propofol, midazolam), targeted temperature management, and metabolic derangements—that shift spectral content in directions that mimic both ageing and de-ageing; and (iii) the *premorbid neurobiological trait* of interest (chronological age and its neural correlates). In healthy recordings, the injury term is zero and the state term is small, so the age component dominates. In post-cardiac-arrest coma, the injury term overwhelms the others, collapsing inter-individual age differences into a narrow residual.

The state term is not merely additive noise. Under standard TTM protocol, therapeutic hypothermia (33–36°C) combined with deep sedation (propofol/midazolam, often with chemical paralysis) during the first $\sim$24 hours actively compresses the EEG into a low-frequency, low-complexity regime that is largely independent of both age and injury severity. This iatrogenic variance compression is distinct from the injury term: even a neurologically intact 25-year-old brain subjected to the same pharmacological–thermal regime would produce EEG indistinguishable from that of a severely injured 80-year-old under the same conditions. Because I-CARE recordings span the first 72 hours, a substantial proportion of the data was acquired during this "deep freeze" window. Any model trained on these recordings must extract age signal from the post-rewarming fraction alone, diluted by pharmacologically homogenised segments.

A specific concern is whether CN-BAG might index sedation depth rather than age-related neurophysiology. Propofol enhances GABAergic inhibition, slowing the decay

time of synaptic potentials, inducing anteriorisation of alpha rhythms, and steepening the aperiodic ($1/f$) slope [7]—precisely the features on which brain-age models rely. Pharmacological and age-related spectral changes are therefore not merely correlated but neurophysiologically convergent, making their separation from scalp EEG alone fundamentally difficult. Without dose-level metadata, the contribution of sedation to the compound signal cannot be quantified. However, the coma-normative design provides a structural defence: because the reference population shares the ICU sedation regime, shared pharmacological effects are absorbed into the population baseline rather than appearing as inter-individual variance. CN-BAG therefore reflects *deviations from the sedated-coma norm*, not sedation per se.

The definitive Phase 2 result (EXP-201: MAE $= 12.21$ y; EXP-202 sealed test: MAE $= 12.90$ y) confirms the null across both paradigms. Notably, sex classification served as a diagnostic indicator throughout: the weak Phase 1 signal (best cross-validated AUC $= 0.600$, permutation $p = 0.001$, barely above chance) and complete Phase 2 degeneration (BAcc $= 0.500$) from the comatose arrays jointly implied a priori that individual-level age trait variance was likely also obliterated, since sex is a coarser and more robust biometric target. The CPC 1–4 sensitivity run (EXP-201-excpc5: MAE $= 9.87 \pm 1.85$ y, below the CPC 1–4 subgroup null of 10.89 y at the point estimate) provides a directional signal consistent with improved age estimation once the most severely injured patients are excluded. However, this sensitivity run was designed with a sixfold larger per-patient window budget (300 vs. 50); the primary experiments (EXP-201/202) could not be rerun at this increased budget before submission. The improvement at the point estimate is therefore most plausibly attributable to the larger training window pool rather than to outcome-group recomposition, as the window count constitutes the primary experimental difference between the two runs.

The broader implication is that any EEG biomarker targeting individual-level traits (brain age, cognitive reserve, neurodegenerative staging) will face the same inherent limit when applied to populations with severe acute injury. The boundary between recoverable trait information and overwhelming acute perturbation is not a fixed property of the measurement instrument but depends on the severity of the biological context.

### 6.4.1 WLST Label Confounding

CPC 5 patients (58.2% of the cohort) include deaths from withdrawal of life-sustaining therapy (WLST), which is itself a clinical decision influenced by EEG findings [5]. This

creates a circular confound for any model trained to predict CPC: the model may learn features that correlate with the WLST decision rather than the patient's biological trajectory.

The CN-BAG design addresses this at two levels. First, the primary prediction target is chronological age—a label fixed at birth that precedes and is structurally independent of all clinical decisions. No WLST decision can change a patient's age, so the training signal is uncontaminated. Second, the dual analysis protocol evaluates CN-BAG–outcome associations both with and without CPC 5 patients.

Because the corrected CN-BAG shows no primary outcome association in the full cohort ($p = 0.058$), while the WLST-excluded sensitivity analysis yields a nominally significant signal ($p = 0.009$), the dual protocol proves its value: the WLST confound is not merely theoretical but actively modulates outcome-association results. However, given the underpowered design ($n_{\text{poor}} = 20$), this secondary signal is best characterised as hypothesis-generating evidence of a residual prognostic signal rather than proof of a real association.

The ethical concern is structural and persists independently of any single study's findings: training on WLST-biased outcome data reiterates the withdrawal decision as if it were biological ground truth, and deploying such a model clinically risks reinforcing the very bias it inherits. The dual analysis protocol remains essential for any future iteration that achieves a detectable CN-BAG signal.

### 6.4.2 WLST-Reduced Sensitivity Signal

The clearest positive result in this thesis is a nominal sensitivity signal after excluding CPC 5 patients. The cohort's age and outcome distribution is summarised in Figure 3.7. By shifting the outcome boundary to remove patients in whom recorded death may follow a clinical care-withdrawal decision, a nominally significant bias-corrected CN-BAG difference is identified between good (CPC 1–2) and poor neurological outcome among survivors (CPC 3–4: $p = 0.009$, AUROC $= 0.680$). This supports the Coma-Normative framework, but it must be qualified: the poor-outcome group is severely underpowered ($n_{\text{poor}} = 20$, power $= 57.6\%$), excluding CPC 5 reduces but does not eliminate WLST confounding, and the primary full-cohort analysis does not reach significance ($p = 0.058$).[1] The result is therefore hypothesis-generating rather than confirmatory: the relative

[1] The sensitivity run differed from the primary experiments in two parameters simultaneously: CPC composition (CPC 1–4 versus the full cohort) and window count per patient (300 versus 50). The improvement cannot be attributed to either variable in isolation.

deviation from the population norm may still be informative even when absolute age prediction fails. Accordingly, the WLST-aware analysis should be read as a bias-control framework for research inference, not as a direct bedside replacement for established multimodal prognostic pathways.

The comparison between raw and bias-corrected BAG in `EXP-203` (Section 5.6.1) provides a critical lesson in methodological rigour. The raw BAG yielded a spurious correlation with outcome ($p = 0.0019$), driven by chronological-age imbalance rather than a biologically meaningful effect. This validates the population-relative normalisation guard and the post-hoc bias correction.

The severity of the present null result becomes clearer when placed alongside the performance of comparable methods operating in lower-pathology recording conditions.

## 6.5 Positioning Against Prior Work

The present null result sits at the far end of the spectrum relative to prior brain-age studies. Three reference points frame the contrast.

**Healthy awake benchmark.** Engemann et al. [67] provide the definitive upper-bound benchmark for this class of model: a multi-cohort study ($N > 2{,}500$ healthy adults; Cam-CAN, LEMON, CHBP, TUAB) benchmarking random forests, Riemannian filterbanks, and convolutional networks on awake resting-state M/EEG, achieving MAE $=$ $6.2$–$7.75$ years with $R^2$ up to $0.74$. That baseline assumes an intact, aroused brain in a stable neurophysiological state. In the present cohort, the same class of model collapses into a constant-mean predictor ($R^2 \approx 0$), confirming that the domain shift from healthy-resting to acute-comatose removes essentially all recoverable age variance.

**Sedation as a partial analogue.** Sabbagh et al. [68] study a bridging regime: elective surgical patients under general anaesthesia (propofol or sevoflurane), where the brain is pharmacologically suppressed but structurally intact. Under stable propofol, models achieve MAE $= 8.2$ years ($R^2 = 0.65$)—already worse than the healthy baseline. Crucially, when those same models are transferred to sevoflurane recordings, performance degrades severely to MAE $\approx 15$ years, demonstrating marked sensitivity to pharmacological domain shift. Post-cardiac-arrest coma represents a far more severe version of this challenge: the ICU EEG is shaped simultaneously by deep sedation, targeted temperature management, evolving hypoxic-ischaemic injury, and multi-site monitoring heterogeneity—each a potential source of the domain compression that eliminates age structure.

**Artefact contamination of EEG biomarkers.** Bomatter et al. [27] study 2,494 recordings from TUAB (v3.0.0, ages 0–95) and TDBRAIN, targeting age and sex prediction under varying preprocessing regimes ($R^2 \approx 0.4$–$0.6$ on TUAB, $\approx 0.8$ on TDBRAIN, with autoreject preprocessing; see Table 2.4 and footnote g therein for pipeline dependence). Their central finding is a paradox: applying ICA to remove physiological peripheral artefacts (cardiac, ocular, muscular) consistently *decreases* prediction performance, directly demonstrating that models exploit non-brain signals to reach their reported accuracy. In other words, the performance of standard EEG brain-age pipelines is in part a peripheral biomarker result, not a cortical one. The preprocessing approach employed here was motivated by exactly this concern: the ICU environment saturates EEG with ventilator, cardiac-monitor, and movement artefacts orders of magnitude more severe than those encountered in resting-state laboratory datasets. The Stage 1 hardware-fault exclusion analysis (Section 5.2.1) documents that the domain-adapted pipeline operates on substantially fewer hardware-failure windows than the raw baseline, reducing one obvious confound; however, Stage 1 passage does not certify freedom from physiological artefacts. The null result is therefore most parsimoniously attributed to genuine signal absence rather than residual technical noise—a conclusion supported independently by the MAE invariance across all preprocessing profiles.

Taken together, the three reference points locate the present result as the consequence of a domain shift that progressively destroys age structure: healthy-awake (MAE $\approx 6$ y) $\rightarrow$ anaesthetised-surgical (MAE $\approx 8$–$15$ y depending on drug) $\rightarrow$ acute-hypoxic-ischaemic coma (constant predictor). The same modelling family that succeeds in lower-pathology regimes appears to be mismatched to this acute ICU state.

## 6.6 Alternative Explanations for the Null Result

The constant-predictor result is consistent with a family of explanations that are not mutually exclusive. The biological explanations describe why the age signal may not exist in this population; the computational limitations describe why the pipeline may have failed to detect it even if present. Both classes may contribute, and the current data cannot distinguish their relative weights.

1. **Pharmacological–thermal compression.** Standard TTM protocol combined with deep sedation compresses the EEG into a low-frequency, low-complexity regime largely independent of both age and injury severity. The resulting iatrogenic

variance compression is distinct from the injury term and operates at scale across most of the recording duration.

2. **Regression-to-the-mean artefact.** Without bias correction, the raw BAG yielded a spurious correlation ($p = 0.0019$) driven by chronological-age imbalance, not biologically meaningful effects. The bias correction (`EXP-203`) is essential to any future claim of age predictability in this setting.

3. **NMAR imputation artefact.** Specparam convergence failure on suppressed EEG, combined with median imputation, pulls the sickest patients' representations toward the population centroid, reducing precisely the inter-patient variance required to discriminate age groups.

4. **Feature insufficiency.** Connectivity features (coherence, phase-lag index, graph metrics) were omitted due to prohibitive computation time. These capture inter-regional communication patterns that may carry age-related information not present in univariate spectral summaries. Temporal trend features may also have been extracted from sub-optimal time windows.

5. **Multi-site hardware heterogeneity.** Seven hospitals contributed recordings with different amplifier hardware, electrode types, impedance characteristics, and monitoring protocols. Site-specific variance in amplitude scaling, filter settings, and channel configurations may introduce systematic shifts that mask or mimic age-related patterns.

6. **Acute versus chronic signal regime.** Brain-age estimation has been validated primarily in healthy, resting-state, or chronically ill populations where the EEG reflects a relatively stable neurophysiological state. Acute post-cardiac-arrest coma is qualitatively different: the EEG evolves rapidly over hours to days, dominated by injury patterns (suppression, burst-suppression, periodic discharges) that have no analogue in the populations where brain-age models were originally validated.

7. **Feature vector collision: ageing and acute injury drive anti-collinear spectral trajectories.** Healthy brain ageing is characterised by a gradual *flattening* of the aperiodic $1/f$ slope (reduction in spectral exponent), a modest relative increase in theta/delta activity, and a decline in peak alpha frequency [23, 24]. Severe HIE moves the EEG in the *opposite* direction in the dimensions most diagnostic to brain-age models: TTM and deep sedation actively *steepen* the aperiodic slope [7, 8],

and widespread high-amplitude delta mimics the spectral signature of early childhood or deep sleep—not senescence [6, 69]. Young brains carry characteristically steep aperiodic slopes; acute comatose EEG artificially recreates this juvenile signature. Consequently, a regressor trained on healthy norms predicts the severely injured 70-year-old as *younger* than their true age—the Rejuvenation Artefact (Section 2.3.2)—not older as would follow from collinearity. The trajectories are anti-collinear: the pathological spectral shift actively inverts the age signal, driving the BAG toward zero or reversing its sign, and the severity of the inversion scales with injury severity—explaining the partial signal recovery seen in the CPC 1–4 sensitivity analysis.

8. **Entropy floor effect: complexity collapse eliminates age variance.** Biological ageing is associated with a gradual decline in physiological complexity (sample entropy, multiscale entropy, Lempel–Ziv complexity) [88]—approximately 1–2% per decade in healthy populations. In acute post-cardiac-arrest coma, the anoxic event crushes signal entropy toward the physiological floor (suppression, periodic patterns, isoelectric activity). Once complexity is clipped at this lower boundary, the micro-variance required to discriminate between a 30-year-old and a 70-year-old is obliterated. This entropy floor compounds the feature vector anti-collinearity: injury so thoroughly saturates the spectral dimensions that no residual age-related variance is preserved, leaving zero dynamic range for age discrimination.

9. **Aperiodic cardiac confound in Phase 1 features.** Recent evidence demonstrates that age-related changes in the cortical aperiodic slope are substantially confounded by cardiac volume conduction [9]. Phase 1 used Profile B, which omits ECG-ICA in favour of bipolar-montage passive rejection (Section 3.3.3). The aperiodic features extracted under Profile B therefore carry an unknown cardiac contribution that may have injected noise or, more insidiously, introduced a spurious age-like signal of cardiovascular rather than neural origin. In Phase 2, Profile C retains ECG by design, whereas Profile C-noecg adds ECG-ICA with a Picard correlation threshold of $|r| > 0.90$ as the explicit ablation. The observation that the ECG-removal ablation does not improve on the retained-ECG baseline is reassuring: if accuracy had been materially inflated by cardiac leakage, the ablated run should have changed more strongly. Nevertheless, the Schmidt et al. finding implies that aperiodic-slope features in any brain-age study lacking explicit cardiac-confound quantification should be interpreted with caution; the EECI (Section 3.3.8) was

designed specifically to address this gap.

## 6.7 Methodological Limitations

A primary methodological strength is the temporal scale of the analysis. Whereas standard EEG brain-age benchmarks rely on isolated 20- to 30-minute resting-state recordings [6, 67], this study extracts thousands of statistics per patient across 36 continuous monitoring hours (Section 4.4), covering the full rewarming and sedation-weaning trajectory rather than a single transient snapshot, as motivated by the 89.0% feature-drift rate documented in Section 4.4.1. The remaining subsections address the principal limitations.

### 6.7.1 Missing Expert Annotations

A central limitation is the inability to validate physiological artefact detectors against a verified ground truth. As outlined in the *quantify-not-remove* philosophy (Section 3.3), physiological noise annotations (e.g., ECG, EMG) were structurally dropped from the predictive feature matrix. While this preserved diagnostic patterns, it also means that some pathologies may still share spectral signatures with technical artefacts. Future work should add expert-in-the-loop annotations to calibrate those overlaps. A related design limitation is the fixed upper clamp of $Q_{\max} = 80$ in the prominence-gated notch filter: because the 4-second Welch window cannot resolve peaks narrower than $\approx 0.25$ Hz, the Q adaptation is resolution-limited for narrow mains peaks and the clamp value was not systematically optimised. A wider Welch window or an FIR-based width estimator could yield more precise Q adaptation in future work.

### 6.7.2 Structural Missingness and Imputation

The use of spectral parameterisation (specparam) introduces a systematic bias: models failed to converge on the most severely suppressed EEG recordings. Median imputation pulled these patients toward the cohort mean, which may flatten the inter-patient variance required to discriminate age groups. A compounding structural gap is the near-50% missingness of time-to-ROSC in the I-CARE metadata: because anoxia duration is the primary determinant of hypoxic-ischaemic injury, covariate-adjusted outcome models cannot fully disentangle whether an elevated CN-BAG reflects premorbid neurobiological aging or serves as a proxy for unmeasured anoxia exposure.

### 6.7.3 Normalisation and Non-Stationarity

A design-level limitation of the Phase 1 feature extraction approach is the absence of within-patient amplitude normalisation. Phase 2 waveforms are robustly normalised at the signal level (median/IQR per channel, applied during preprocessing), but neither phase has any mechanism to account for the large temporal evolution present in virtually all features.

**Amplitude non-normalisation.** Absolute spectral power features carry electrode–tissue impedance, skull-thickness, and inter-site amplifier scaling as systematic additive offsets. Because these offsets are patient-specific and partly site-specific, they inflate inter-patient variance in ways that are biologically uninformative. Standard healthy-population brain-age benchmarks [67] mitigate this via robust z-scoring against a reference population; the present work omits this step. ElasticNet and Ridge regression, which form the backbone of Phase 1, are sensitive to feature scale: the regularisation penalty is applied uniformly across all coefficients, so high-variance amplitude features are effectively under-penalised relative to unitless ratio features. This scale imbalance may have suppressed the contribution of the most diagnostically relevant features or, conversely, amplified noise from impedance-sensitive channels.

**Temporal non-stationarity.** Exploratory analysis confirmed that 89.0% of extracted features show statistically significant monotonic trends across the first 36 hours of recording (Figure 4.9), driven by the rewarming ramp of TTM, sedation weaning, and evolving cortical recovery. The patient-level aggregation partially addresses this by computing trajectory statistics (slope, IQR, skewness) alongside point estimates, but it does not explicitly normalise for the temporal position of each window within the patient trajectory, nor for heterogeneity in total recording duration. Patients with short recordings are represented predominantly by the acute sedated phase; patients with long recordings additionally cover the recovery window. These two subpopulations are not directly comparable in the aggregated feature space, and the models have no mechanism to condition on recording duration or epoch position. Future work should explore time-relative normalisation—for example expressing each window's features as deviations from the patient's own baseline window at hour zero—or explicit trajectory modelling, such as representing each patient as a time-indexed functional curve rather than a bag of statistics. A further consequence is a potential shift in cohort composition: an exploratory quality sweep (EXP-313) found that at the 5% per-channel exclusion threshold, TTM patients

were statistically over-represented among those excluded ($p = 0.026$), indicating that the automated pipeline partially misclassified therapeutic hypothermia's EEG suppression as technical artefact; the conservative thresholds adopted in the final calibration were chosen partly to mitigate this, but residual demographic skew in the excluded fraction cannot be excluded.

### 6.7.4 Model Capacity and Temporal Context

Phase 2 deep learning was trained with a 300-window-per-patient cap. While the definitive run (EXP-201) indicates that the null result is not simply a small-$N$ artefact, there is still a distinction between total data volume and *temporal context* ($T$). Sequence models like the CNN-BiLSTM-Attention may need substantially more continuous context to distinguish subtle *thalamocortical recovery* transitions from the non-stationarity of the HIE state. Future work should therefore test the full multi-day trajectory uncompressed, rather than assuming that isolated windows are enough.

A second limitation concerns cohort scale for deep-learning inference. Although $N_{\text{usable}} = 604$ is substantial for ICU EEG, it remains far below the population sizes typically used to stabilise high-capacity biometric models in less pathological regimes. A compounding factor specific to Phase 2 is that the quality filter reduced the effective deep-learning cohort from the full 483-patient development set to the 286 patients who contributed at least one valid 60 s window; this restriction only became apparent when the dataset loader reported fewer patients than the split file at evaluation time, and the full 483-patient run would need to be repeated to confirm whether the excluded patients alter the result. The critical counter-evidence is architectural convergence: Phase 1 classical ML (XGBoost) — which is generally less data-hungry than high-capacity deep learning, though a 1,578-feature space at $N = 604$ may still impose some constraints even for tree-based methods — produces an identical constant-mean null result across all five preprocessing profiles. That two architecturally independent pipelines converge on the same floor implicates a biological rather than statistical explanation. The training curves for all four Phase 2 architectures (Figure 5.7) confirm that this convergence is not the result of failed optimisation: all models reduce their training loss by approximately an order of magnitude across 30 epochs, yet their validation performance converges to the same near-null plateau. Accordingly, this thesis cannot claim that all model families at larger scale would necessarily fail; rather, it shows that under a rigorously controlled, clinically realistic ICU setting, both the feature-based and end-to-end pipelines tested

here converge to near-null age recovery.

### 6.7.5 SWOT Analysis

**Table 6.1:** Structured assessment of the CN-BAG methodology across four dimensions.

| | |
|---|---|
| **Strengths** | Decision-independent label (chronological age) is structurally immune to WLST contamination; coma-normative training avoids the Rejuvenation Artefact; compound-biomarker framing treats State–Trait Inseparability as a finding rather than a flaw; hardware-physics preprocessing thresholds are generalisable across acquisition sites. |
| **Weaknesses** | No sedation or drug-dosing metadata; the 36-hour window captures predominantly the pharmacological phase, limiting conclusions about underlying neuronal integrity; single-dataset external validity untested; sample size constrains deep learning capacity. |
| **Opportunities** | Serial CN-BAG trajectories over the ICU stay as a continuous recovery biomarker [83]; complexity-informed extensions integrating non-linear measures (Lempel-Ziv complexity, transfer entropy) [71]; pharmacological metadata integration as an explicit statistical covariate once drug-dosing records become available; multi-centre external validation to establish site-independence. |
| **Threats** | Pharmacological confounds (propofol, TTM) induce spectral shifts that directly overlap with age-sensitive bands [11]; cardiac volume conduction as a latent confound if EECI calibration is not completed (Section 3.3.8); State–Trait Inseparability may render the BAG insensitive to longitudinal changes if the acute state fully dominates the signal [6]; WLST-driven label noise in any downstream outcome correlation. |

## 6.8 Future Work

State-trait inseparability identifies a plausible biological boundary for computational neuroprognostication. Future work should target:

- **Uncompressed Multimodal Fusion:** Ingesting synchronous clinical covariates (sedation doses, GCS) to disambiguate pharmacological states from injury.
- **The Jaccard Shield Analysis:** Using feature-overlap metrics to map the intersection between age-predictive and outcome-predictive features, providing a formal check on state-trait contamination.

- **Long-Range Sequence Modelling:** Utilising distributed training to model multi-day trajectories without temporal downsampling.

- **Temporal-Phase Stratification:** Evaluating models separately on windows from the acute sedated phase and from the post-rewarming recovery window, in the subset of patients with sufficiently long recordings. The current analysis aggregates over the full available recording duration (median 46.1 h, but with substantial variation across patients), which dilutes any recovery-phase signal with pharmacologically homogenised acute-phase data. Stratified evaluation on the longer-recording subset would directly probe whether the state-trait eclipse weakens once sedation and temperature management have been withdrawn—an analysis deferred due to time and resource constraints.

- **Zero-Shot Transfer:** Quantifying the *Rejuvenation Artefact* by applying healthy-normed models to the I-CARE corpus and measuring the resulting domain shift.

- **Expert Annotation Pipeline:** Establishing a neurophysiologist-in-the-loop annotation framework to distinguish clinical myoclonus from technical muscle artefact, enabling the quality mask to preserve rather than conflate pathological signal.

- **EECI Calibration and Validation (RO-4):** The *EEG ECG Contamination Index* is proposed here as a novel multi-domain metric for quantifying cardiac contamination in clinical EEG, combining cross-correlation, dual-band spectral coherence, Phase-Locking Value, harmonic coherence, R-peak-locked RMS, and Mutual Information into a per-channel, per-window severity score (see Sections B.2.4 and 3.3.8 for the full formula and component rationale). To our knowledge, no such composite index exists in the literature. The prototype is implemented; the outstanding step is empirical calibration of the component weights via neurophysiologist-annotated ground truth. Once calibrated, the EECI would function as a standalone reproducibility safeguard for brain age and other EEG biomarker studies in any setting where cardiac volume conduction is uncontrolled.

- **Longitudinal CN-BAG Trajectories:** Treating CN-BAG as a time series across the 72-hour monitoring window rather than a scalar, to test whether the direction of brain age change—improving versus deteriorating—carries prognostic signal beyond the initial level.

The contributions of this thesis—the Coma-Normative framework, the hardware-physics bitmasking strategy, the dual-analysis protocol, and the EECI proposal—provide a foundation for future researchers working with the severely disrupted signal environment of comatose EEG.

Chapter 7 summarises these contributions and provides the synthesis.

# Chapter 7

# Conclusion

This thesis set out to test whether a brain age gap metric derived from post-cardiac-arrest EEG could serve as a prognostic biomarker that avoids the WLST self-fulfilling prophecy inherent in outcome-trained classifiers. The primary empirical hypothesis was not supported: no detectable age signal was recoverable from the feature-based representation of comatose ICU EEG ($R^2 \approx 0$; Phase 1 cross-validated MAE $\approx 12.2$ years; Phase 2 sealed-test MAE $= 12.90$ years, 95 % CI [10.73, 15.24]). The contributions are summarised below from conceptual framework to implementation infrastructure.

**The CN-BAG framework.** The Coma-Normative Brain Age Gap defines brain age relative to the coma population itself, bypassing the Rejuvenation Artefact that would invalidate healthy-normative models applied to severely suppressed EEG. The framework is applicable to any clinical population in which the target condition disrupts the normal age–EEG relationship, and its OLS bias-correction step was empirically validated: the spurious raw-BAG association ($p = 0.0019$, driven by age-outcome collinearity) disappeared after correction, confirming that this bias-correction step was necessary. A WLST-excluded sensitivity analysis (CPC 1–4, $n = 162$) yielded a nominally significant CN-BAG–outcome association ($p = 0.009$; AUROC $= 0.680$ [0.555, 0.801]), providing a first grounding for the coma-normative framework under reduced outcome-label confounding, albeit in an underpowered subgroup.

**State-invariant preprocessing architecture.** The two-stage quality architecture—hardware fault masking with physics-grounded absolute thresholds, followed by non-destructive neurological pattern annotation—was designed specifically for the signal conditions of ICU coma EEG, where standard pipelines calibrated to healthy awake

recordings misclassify pathological suppression as artefact. It preserves burst-suppression and other pathological patterns while rejecting amplifier artefacts, and operates identically regardless of background EEG state, achieving a median clean ratio of 0.968 across 604 patients.

**Preprocessing impact assessment.** The profile comparison (Profiles A, B, and E) provides a systematic evaluation of how preprocessing choices affect brain age prediction in coma; Profiles C and D were not evaluated at the feature level. The finding that raw and fully preprocessed profiles yield statistically indistinguishable age MAE (Wilcoxon $p_{\text{adj}} = 1.000$; Section 5.2) suggests that the limiting factor is likely neurophysiological rather than technical, though the possibility that the first-36-hour recording window captures predominantly pharmacologically suppressed signal—or that cleaning was insufficient—cannot be excluded.

**State–trait inseparability.** In severe HIE, irreversible neuronal injury fuses acute state with premorbid neural trait, collapsing inter-individual age differences into a narrow residual. The intra-patient memorisation experiment (Sex BAcc $= 0.988$, Age MAE $= 10.13$ years; Section 5.4) confirms that the pipeline captured individual biometric fingerprints but could not generalise them across patients: the failure is biological, not technical.

**Experimental infrastructure.** The HPC Load-Once/Fork orchestration, feature-adaptive aggregation scheme, and dual-workflow quality mask are directly reusable for future ICU EEG research. The declarative profile system ensures reproducibility and cross-study comparability. EECI development and population-level validation (RO-4) are deferred to future work; the computation code is integrated into the preprocessing pipeline.

Post-cardiac-arrest prognostication operates under a fundamental constraint: every prediction that influences a withdrawal decision risks becoming self-fulfilling. This thesis proposed an alternative—predicting a label that cannot be changed by clinical decisions—and tested whether it works. For feature-based and end-to-end deep learning on the I-CARE cohort, the answer is that it does not, at least not at the current scale and with the current signal representation.

This negative result is scientifically informative. It characterises the state–trait boundary in severe brain injury, demonstrates that preprocessing is not the bottleneck, and

provides a complete experimental framework for future attempts. The preprocessing design, CN-BAG formalism, and dual-analysis WLST protocol are ready for reuse in studies with larger cohorts, pharmacological metadata, or sequence-modelling architectures. This work does not claim that comatose EEG contains no age-related information—only that the current feature extraction and modelling approach cannot recover it. Whether temporal trajectory modelling, self-supervised representations, or multimodal fusion can succeed where hand-crafted features failed, remains an open question.

Ultimately, this work suggests that in some domains, method choice becomes meaningful only after the context is properly understood; without that grounding, the modelling task itself remains epistemologically flawed. The value of this thesis is therefore methodological and scientific as much as it is ethical when it defines a grounded negative result, a cautious positive sensitivity result, and a reproducible pipeline for future work, without claiming definitive answers.

# References


[1] Lise Witten, Ryan Gardner, Mathias J. Holmberg, et al. “Reasons for Death in Patients Successfully Resuscitated from Out-of-Hospital and in-Hospital Cardiac Arrest”. In: *Resuscitation* 136 (Mar. 2019), pp. 93–99. ISSN: 03009572. DOI: 10.1016/j.resuscitation.2019.01.031.

[2] Claudio Sandroni, Sonia D’Arrigo, Sofia Cacciola, et al. “Prediction of Poor Neurological Outcome in Comatose Survivors of Cardiac Arrest: A Systematic Review”. In: *Intensive Care Medicine* 46.10 (Oct. 1, 2020), pp. 1803–1851. ISSN: 1432-1238. DOI: 10.1007/s00134-020-06198-w.

[3] Jonathan Elmer, Cesar Torres, Tom P. Aufderheide, et al. “Association of Early Withdrawal of Life-Sustaining Therapy for Perceived Neurological Prognosis with Mortality after Cardiac Arrest”. In: *Resuscitation* 102 (May 2016), pp. 127–135. ISSN: 03009572. DOI: 10.1016/j.resuscitation.2016.01.016.

[4] Irina Dragancea, Matthew P. Wise, Nawaf Al-Subaie, et al. “Protocol-Driven Neurological Prognostication and Withdrawal of Life-Sustaining Therapy after Cardiac Arrest and Targeted Temperature Management”. In: *Resuscitation* 117 (Aug. 1, 2017), pp. 50–57. ISSN: 0300-9572, 1873-1570. DOI: 10.1016/j.resuscitation.2017.05.014. PMID: 28506865.

[5] Romergryko G. Geocadin, Clifton W. Callaway, Ericka L. Fink, et al. “Standards for Studies of Neurological Prognostication in Comatose Survivors of Cardiac Arrest: A Scientific Statement From the American Heart Association”. In: *Circulation* 140.9 (Aug. 27, 2019). ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIR.0000000000000702.

[6] Lukas A.W. Gemein, Robin T. Schirrmeister, Joschka Boedecker, and Tonio Ball. “Brain Age Revisited: Investigating the State vs. Trait Hypotheses of EEG-derived Brain-Age Dynamics with Deep Learning”. In: *Imaging Neuroscience* 2 (July 8,

2024), imag-2–00210. ISSN: 2837-6056. DOI: 10.1162/imag_a_00210. PMID: 40800431.

[7] Emery N. Brown, Ralph Lydic, and Nicholas D. Schiff. "General Anesthesia, Sleep, and Coma". In: *New England Journal of Medicine* 363.27 (Dec. 30, 2010), pp. 2638–2650. ISSN: 0028-4793. DOI: 10.1056/NEJMra0808281.

[8] Songyu Chen, Brittany Bolduc Lachance, Liang Gao, and Xiaofeng Jia. "Targeted Temperature Management and Early Neuro-Prognostication after Cardiac Arrest". In: *Journal of Cerebral Blood Flow & Metabolism* 41.6 (June 2021), pp. 1193–1209. ISSN: 0271-678X, 1559-7016. DOI: 10.1177/0271678X20970059.

[9] Fabian Schmidt, Sarah K Danböck, Eugen Trinka, Dominic P Klein, Gianpaolo Demarchi, and Nathan Weisz. "Age-Related Changes in 'Cortical' 1/f Dynamics Are Linked to Cardiac Activity". In: *eLife* 13 (Oct. 2, 2025). Ed. by Björn Herrmann and Huan Luo, RP100605. ISSN: 2050-084X. DOI: 10.7554/eLife.100605.

[10] Edilberto Amorim, Wei-Long Zheng, Mohammad M. Ghassemi, et al. "The International Cardiac Arrest Research Consortium Electroencephalography Database". In: *Critical Care Medicine* 51.12 (Dec. 1, 2023), pp. 1802–1811. ISSN: 1530-0293. DOI: 10.1097/CCM.0000000000006074. PMID: 37855659.

[11] P. L. Purdon, K. J. Pavone, O. Akeju, et al. "The Ageing Brain: Age-dependent Changes in the Electroencephalogram during Propofol and Sevoflurane General Anaesthesia". In: *British Journal of Anaesthesia* 115 (Suppl 1 July 1, 2015), pp. i46–i57. ISSN: 0007-0912. DOI: 10.1093/bja/aev213. PMID: 26174300.

[12] Jose Antonio Urigüen and Begoña Garcia-Zapirain. "EEG Artifact Removal—State-of-the-Art and Guidelines". In: *Journal of Neural Engineering* 12.3 (Apr. 2015), p. 031001. ISSN: 1741-2552. DOI: 10.1088/1741-2560/12/3/031001.

[13] Morteza Zabihi, Alireza Chaman Zar, Pulkit Grover, and Eric S. Rosenthal. "HyperEnsemble Learning from Multimodal Biosignals to Robustly Predict Functional Outcome after Cardiac Arrest". In: 2023 Computing in Cardiology Conference. Nov. 26, 2023. DOI: 10.22489/CinC.2023.142.

[14] Marleen C. Tjepkema-Cloostermans, Catarina da Silva Lourenço, Barry J. Ruijter, et al. "Outcome Prediction in Postanoxic Coma With Deep Learning*". In: *Critical Care Medicine* 47.10 (Oct. 2019), p. 1424. ISSN: 1530-0293. DOI: 10.1097/CCM.0000000000003854.

[15] Frédéric Zubler, Andreas Steimer, Rebekka Kurmann, et al. "EEG Synchronization Measures Are Early Outcome Predictors in Comatose Patients after Cardiac Arrest". In: *Clinical Neurophysiology* 128.4 (Apr. 1, 2017), pp. 635–642. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2017.01.020.

[16] Andrea O. Rossetti, Mauro Oddo, Giancarlo Logroscino, and Peter W. Kaplan. "Prognostication after Cardiac Arrest and Hypothermia: A Prospective Study". In: *Annals of Neurology* 67.3 (Mar. 2010), pp. 301–307. ISSN: 1531-8249. DOI: 10.1002/ana.21984. PMID: 20373341.

[17] Aatif M. Husain. "Electroencephalographic Assessment of Coma". In: *Journal of Clinical Neurophysiology* 23.3 (June 2006), p. 208. ISSN: 0736-0258. DOI: 10.1097/01.wnp.0000220094.60482.b5.

[18] F. Lopes da Silva. "Neural Mechanisms Underlying Brain Waves: From Neural Membranes to Networks". In: *Electroencephalography and Clinical Neurophysiology* 79.2 (Aug. 1991), pp. 81–93. ISSN: 0013-4694. DOI: 10.1016/0013-4694(91)90044-5. PMID: 1713832.

[19] Stephan Kratzer, Michael Schneider, David P. Obert, Gerhard Schneider, Paul S. García, and Matthias Kreuzer. "Age-Related EEG Features of Bursting Activity During Anesthetic-Induced Burst Suppression". In: *Frontiers in Systems Neuroscience* 14 (Dec. 3, 2020). ISSN: 1662-5137. DOI: 10.3389/fnsys.2020.599962.

[20] Michael M. Schartner, Robin L. Carhart-Harris, Adam B. Barrett, Anil K. Seth, and Suresh D. Muthukumaraswamy. "Increased Spontaneous MEG Signal Diversity for Psychoactive Doses of Ketamine, LSD and Psilocybin". In: *Scientific Reports* 7 (Apr. 19, 2017), p. 46421. ISSN: 2045-2322. DOI: 10.1038/srep46421. PMID: 28422113.

[21] Marvic Joseph S. Amoranto and Marc Laurence L. Fernandez. "Unreactive Alpha or Alpha-Theta Coma Electroencephalogram Patterns with Favorable Neurologic Outcomes: A Report of Three Cases and Literature Review". In: *Journal of Neurocritical Care* 17.2 (Nov. 14, 2024), pp. 67–74. ISSN: 2508-1349. DOI: 10.18700/jnc.240020.

[22] Claudio Babiloni, Robert J. Barry, Erol Başar, et al. "International Federation of Clinical Neurophysiology (IFCN) – EEG Research Workgroup: Recommendations on Frequency and Topographic Analysis of Resting State EEG Rhythms. Part 1:

Applications in Clinical Research Studies". In: *Clinical Neurophysiology* 131.1 (Jan. 1, 2020), pp. 285–307. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2019.06.234.

[23] Bradley Voytek, Mark A. Kramer, John Case, et al. "Age-Related Changes in 1/f Neural Electrophysiological Noise". In: *Journal of Neuroscience* 35.38 (Sept. 23, 2015), pp. 13257–13265. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.2332-14.2015. PMID: 26400953.

[24] Thomas Donoghue, Matar Haller, Erik J. Peterson, et al. "Parameterizing Neural Power Spectra into Periodic and Aperiodic Components". In: *Nature Neuroscience* 23.12 (Dec. 2020), pp. 1655–1665. ISSN: 1546-1726. DOI: 10.1038/s41593-020-00744-x. PMID: 33230329.

[25] Yeonsu Kim, Jiho Park, Woosuk Chung, et al. "Age-Related Electroencephalographic Delta and Alpha Oscillations During Sedation with Target-Controlled Propofol Infusion". In: *Journal of Clinical Medicine* 14.9 (Jan. 2025), p. 3024. ISSN: 2077-0383. DOI: 10.3390/jcm14093024.

[26] Jose Luis Perez Velazquez, Diego Martin Mateos, and Richard Wennberg. "Alpha Rhythm, Alpha Coma, and Entropy in EEG". In: *Journal of Clinical Neurophysiology* 43.1 (May 17, 2022), 10.1097/WNP.0000000000001177. ISSN: 0736-0258. DOI: 10.1097/WNP.0000000000001177.

[27] Philipp Bomatter, Joseph Paillard, Pilar Garces, Jörg Hipp, and Denis-Alexander Engemann. "Machine Learning of Brain-Specific Biomarkers from EEG". In: *eBioMedicine* 106 (Aug. 1, 2024), p. 105259. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2024.105259.

[28] Hualou Liang, Fengyu Zhang, and Xin Niu. "Investigating Systematic Bias in Brain Age Estimation with Application to Post-Traumatic Epilepsy". In: *Human Brain Mapping* 40.11 (2019), pp. 3143–3152. DOI: 10.1002/hbm.24588.

[29] Ann-Marie G. de Lange and James H. Cole. "Multimodal Brain-Age Prediction and Cardiovascular Risk: The Whitehall II MRI Sub-Study". In: *NeuroImage* 222 (2020), p. 117292. DOI: 10.1016/j.neuroimage.2020.117292.

[30] Hyunwoong Kim, Seongbeom Park, Sang Won Seo, et al. "A Novel Deep Learning-Based Brain Age Prediction Framework for Routine Clinical MRI Scans". In: *npj Aging* 11.1 (July 29, 2025), p. 70. ISSN: 2731-6068. DOI: 10.1038/s41514-025-00260-x.

[31] James H. Cole, Robert Leech, David J. Sharp, and Alzheimer's Disease Neuroimaging Initiative. "Prediction of Brain Age Suggests Accelerated Atrophy after Traumatic Brain Injury". In: *Annals of Neurology* 77.4 (Apr. 2015), pp. 571–581. ISSN: 1531-8249. DOI: 10.1002/ana.24367. PMID: 25623048.

[32] Gershon Spitz, Amelia J. Hicks, Caroline Roberts, Christopher C. Rowe, and Jennie Ponsford. "Brain Age in Chronic Traumatic Brain Injury". In: *NeuroImage: Clinical* 35 (Jan. 1, 2022), p. 103039. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2022.103039.

[33] Haoqi Sun, Luis Paixao, Jefferson T. Oliva, et al. "Brain Age from the Electroencephalogram of Sleep". In: *Neurobiology of aging* 74 (Feb. 2019), pp. 112–120. ISSN: 0197-4580. DOI: 10.1016/j.neurobiolaging.2018.10.016. PMID: 30448611.

[34] Nikolaos Koutsouleris, Christos Davatzikos, Stefan Borgwardt, et al. "Accelerated Brain Aging in Schizophrenia and beyond: A Neuroanatomical Marker of Psychiatric Disorders". In: *Schizophrenia Bulletin* 40.5 (Sept. 2014), pp. 1140–1153. ISSN: 1745-1701. DOI: 10.1093/schbul/sbt142. PMID: 24126515.

[35] Igor Nenadić, Maren Dietzek, Kerstin Langbein, Heinrich Sauer, and Christian Gaser. "BrainAGE Score Indicates Accelerated Brain Aging in Schizophrenia, but Not Bipolar Disorder". In: *Psychiatry Research. Neuroimaging* 266 (Aug. 30, 2017), pp. 86–89. ISSN: 1872-7506. DOI: 10.1016/j.pscychresns.2017.05.006. PMID: 28628780.

[36] Heath R. Pardoe, James H. Cole, Karen Blackmon, Thomas Thesen, Ruben Kuzniecky, and Human Epilepsy Project Investigators. "Structural Brain Changes in Medically Refractory Focal Epilepsy Resemble Premature Brain Aging". In: *Epilepsy Research* 133 (July 2017), pp. 28–32. ISSN: 1872-6844. DOI: 10.1016/j.eplepsyres.2017.03.007. PMID: 28410487.

[37] Luis Paixao, Pooja Sikka, Haoqi Sun, et al. "Excess Brain Age in the Sleep Electroencephalogram Predicts Reduced Life Expectancy". In: *Neurobiology of aging* 88 (Apr. 2020), pp. 150–155. ISSN: 0197-4580. DOI: 10.1016/j.neurobiolaging.2019.12.015. PMID: 31932049.

[38] J H Cole, S J Ritchie, M E Bastin, et al. "Brain Age Predicts Mortality". In: *Molecular Psychiatry* 23.5 (May 2018), pp. 1385–1392. ISSN: 1359-4184. DOI: 10.1038/mp.2017.62. PMID: 28439103.

[39] J. S. Barlow. “Methods of Analysis of Nonstationary EEGs, with Emphasis on Segmentation Techniques: A Comparative Review”. In: *Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society* 2.3 (July 1985), pp. 267–304. ISSN: 0736-0258. DOI: 10.1097/00004691-198507000-00005. PMID: 3916847.

[40] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. “The PREP Pipeline: Standardized Preprocessing for Large-Scale EEG Analysis”. In: *Frontiers in Neuroinformatics* 9 (June 18, 2015). ISSN: 1662-5196. DOI: 10.3389/fninf.2015.00016.

[41] Mainak Jas, Denis A. Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. “Autoreject: Automated Artifact Rejection for MEG and EEG Data”. In: *NeuroImage* 159 (Oct. 1, 2017), pp. 417–429. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2017.06.030.

[42] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. “ICLabel: An Automated Electroencephalographic Independent Component Classifier, Dataset, and Website”. In: *NeuroImage* 198 (Sept. 2019), pp. 181–197. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2019.05.026.

[43] M. N. van Stigt, C. Ruiz Camps, J. M. Coutinho, H. A. Marquering, B. S. Doelkahar, and W. V. Potters. “The Effect of Artifact Rejection on the Performance of a Convolutional Neural Network Based Algorithm for Binary EEG Data Classification”. In: *Biomedical Signal Processing and Control* 85 (Aug. 1, 2023), p. 105032. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2023.105032.

[44] Roman Kessler, Alexander Enge, and Michael A. Skeide. “How EEG Preprocessing Shapes Decoding Performance”. In: *Communications Biology* 8.1 (July 10, 2025), p. 1039. ISSN: 2399-3642. DOI: 10.1038/s42003-025-08464-3.

[45] Elnaz Lashgari, Dehua Liang, and Uri Maoz. “Data Augmentation for Deep-Learning-Based Electroencephalography”. In: *Journal of Neuroscience Methods* 346 (Dec. 1, 2020), p. 108885. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2020.108885.

[46] Li Wang, Weijian Huang, Zhao Yang, and Chun Zhang. “Temporal-Spatial-Frequency Depth Extraction of Brain-Computer Interface Based on Mental Tasks”. In: *Biomedical Signal Processing and Control* 58 (Apr. 1, 2020), p. 101845. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2020.101845.

[47] Xiang Zhang and Lina Yao. *Deep Learning for EEG-Based Brain-Computer Interfaces: Representations, Algorithms and Applications*. New Jersey: World Scientific, 2022. ISBN: 978-1-78634-958-3.

[48] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. "Deep Learning for Electroencephalogram (EEG) Classification Tasks: A Review". In: *Journal of Neural Engineering* 16.3 (June 1, 2019), p. 031001. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/ab0ab5.

[49] Dominik Walther, Johannes Viehweg, Jens Haueisen, and Patrick Mäder. "A Systematic Comparison of Deep Learning Methods for EEG Time Series Analysis". In: *Frontiers in Neuroinformatics* 17 (Feb. 23, 2023), p. 1067095. ISSN: 1662-5196. DOI: 10.3389/fninf.2023.1067095.

[50] Jin Xie, Jie Zhang, Jiayao Sun, et al. "A Transformer-Based Approach Combining Deep Learning Network and Spatial-Temporal Information for Raw EEG Classification". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), pp. 2126–2136. ISSN: 1558-0210. DOI: 10.1109/TNSRE.2022.3194600.

[51] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Aug. 13, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. arXiv: 1603.02754 [cs].

[52] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine." In: *The Annals of Statistics* 29.5 (Oct. 2001), pp. 1189–1232. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1013203451.

[53] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. "EEGNet: A Compact Convolutional Neural Network for EEG-based Brain–Computer Interfaces". In: *Journal of Neural Engineering* 15.5 (July 2018), p. 056013. ISSN: 1741-2552. DOI: 10.1088/1741-2552/aace8c.

[54] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*. Version 2. 2018. DOI: 10.48550/ARXIV.1803.01271. Pre-published.

[55] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 4, 2017, pp. 4768–4777. ISBN: 978-1-5108-6096-4.

[56] Chenxi Sun, Jin Jing, Niels Turley, et al. "Harvard Electroencephalography Database: A Comprehensive Clinical Electroencephalographic Resource from Four Boston Hospitals". In: *Epilepsia* 66.9 (Sept. 2025), pp. 3411–3425. ISSN: 0013-9580, 1528-1167. DOI: 10.1111/epi.18487.

[57] S. F. Quan, B. V. Howard, C. Iber, et al. "The Sleep Heart Health Study: Design, Rationale, and Methods". In: *Sleep* 20.12 (Dec. 1997), pp. 1077–1085. ISSN: 0161-8105. PMID: 9493915.

[58] Iyad Obeid and Joseph Picone. "The Temple University Hospital EEG Data Corpus". In: *Frontiers in Neuroscience* 10 (May 13, 2016). ISSN: 1662-453X. DOI: 10.3389/fnins.2016.00196.

[59] Hanneke van Dijk, Guido van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde van Ruth, and Martijn Arns. "The Two Decades Brainclinics Research Archive for Insights in Neurophysiology (TDBRAIN) Database". In: *Scientific Data* 9.1 (June 14, 2022), p. 333. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01409-z.

[60] Anahit Babayan, Miray Erbey, Deniz Kumral, et al. "A Mind-Brain-Body Dataset of MRI, EEG, Cognition, Emotion, and Peripheral Physiology in Young and Old Adults". In: *Scientific Data* 6.1 (Feb. 12, 2019), p. 180308. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.308.

[61] Edilberto Amorim, Wei-Long Zheng, Jong Woo Lee, et al. *I-CARE: International Cardiac Arrest REsearch Consortium Database*. Version 2.0. PhysioNet, 2023. DOI: 10.13026/AVEK-0P97.

[62] Edilberto Amorim, Wei-Long Zheng, Mohammad M. Ghassemi, et al. *The International Cardiac Arrest Research (I-CARE) Consortium Electroencephalography Database*. Aug. 28, 2023. DOI: 10.1101/2023.08.28.23294672. Prepublished.

[63] Sahar Zafar, Tobias Loddenkemper, Jong Woo Lee, et al. *Harvard Electroencephalography Database*. Version 4.1. BDSP. DOI: 10.60508/K85B-FC87.

[64] Cristina Gil Ávila, Felix S. Bott, Laura Tiemann, et al. "DISCOVER-EEG: An Open, Fully Automated EEG Pipeline for Biomarker Discovery in Clinical Neuroscience". In: *Scientific Data* 10.1 (Sept. 11, 2023), p. 613. ISSN: 2052-4463. DOI: 10.1038/s41597-023-02525-0.

[65] Jason R. Taylor, Nitin Williams, Rhodri Cusack, et al. "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) Data Repository: Structural and Functional MRI, MEG, and Cognitive Data from a Cross-Sectional Adult Lifespan Sample". In: *NeuroImage*. Data Sharing Part II 144 (Jan. 1, 2017), pp. 262–269. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2015.09.018.

[66] Di Zhang, Yichong She, Jinbo Sun, et al. "Brain Age Estimation from Overnight Sleep Electroencephalography with Multi-Flow Sequence Learning". In: *Nature and Science of Sleep* Volume 16 (July 2024), pp. 879–896. ISSN: 1179-1608. DOI: 10.2147/NSS.S463495.

[67] Denis A. Engemann, Apolline Mellot, Richard Höchenberger, et al. "A Reusable Benchmark of Brain-Age Prediction from M/EEG Resting-State Signals". In: *NeuroImage* 262 (Nov. 15, 2022), p. 119521. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2022.119521.

[68] David Sabbagh, Jérôme Cartailler, Cyril Touchard, et al. "Repurposing Electroencephalogram Monitoring of General Anaesthesia for Building Biomarkers of Brain Ageing: An Exploratory Study". In: *BJA Open* 7 (Sept. 1, 2023), p. 100145. ISSN: 2772-6096. DOI: 10.1016/j.bjao.2023.100145.

[69] J. Hofmeijer and M. J. A. M. van Putten. "EEG in Postanoxic Coma: Prognostic and Diagnostic Value". In: *Clinical Neurophysiology* 127.4 (Apr. 1, 2016), pp. 2047–2055. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2016.02.002.

[70] Edilberto Amorim, Wei-Long Zheng, Jin Jing, et al. "Neurophysiology State Dynamics Underlying Acute Neurologic Recovery After Cardiac Arrest". In: *Neurology* 101.9 (Aug. 29, 2023), e940–e952. DOI: 10.1212/WNL.0000000000207537.

[71] Sigurd L. Alnes, Marzia De Lucia, Andrea O. Rossetti, and Athina Tzovara. "Complementary Roles of Neural Synchrony and Complexity for Indexing Consciousness and Chances of Surviving in Acute Coma". In: *NeuroImage* 245 (Dec. 15, 2021), p. 118638. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2021.118638.

[72] Jeannette Hofmeijer, Marleen C. Tjepkema-Cloostermans, and Michel J. A. M. van Putten. "Burst-Suppression with Identical Bursts: A Distinct EEG Pattern with Poor Outcome in Postanoxic Coma". In: *Clinical Neurophysiology* 125.5 (May 1, 2014), pp. 947–954. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2013.10.017.

[73] Mohammad M. Ghassemi, Edilberto Amorim, Tuka Alhanai, et al. "Quantitative Electroencephalogram Trends Predict Recovery in Hypoxic-Ischemic Encephalopathy*". In: *Critical Care Medicine* 47.10 (Oct. 2019), p. 1416. ISSN: 1530-0293. DOI: 10.1097/CCM.0000000000003840.

[74] M. M. Admiraal, L. A. Ramos, S. Delgado Olabarriaga, H. A. Marquering, J. Horn, and A. F. van Rootselaar. "Quantitative Analysis of EEG Reactivity for Neurological Prognostication after Cardiac Arrest". In: *Clinical Neurophysiology* 132.9 (Sept. 1, 2021), pp. 2240–2247. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2021.07.004.

[75] Jan Claassen, Kevin Doyle, Adu Matory, et al. "Detection of Brain Activation in Unresponsive Patients with Acute Brain Injury". In: *New England Journal of Medicine* 380.26 (June 27, 2019), pp. 2497–2505. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1812757.

[76] Stanley D. T. Pham, Hanneke M. Keijzer, Barry J. Ruijter, et al. "Outcome Prediction of Postanoxic Coma: A Comparison of Automated Electroencephalography Analysis Methods". In: *Neurocritical Care* 37.2 (Aug. 1, 2022), pp. 248–258. ISSN: 1556-0961. DOI: 10.1007/s12028-022-01449-8.

[77] Arnaud Delorme. "EEG Is Better Left Alone". In: *Scientific Reports* 13.1 (Feb. 9, 2023), p. 2372. ISSN: 2045-2322. DOI: 10.1038/s41598-023-27528-0.

[78] Xin Gao, Kai Gui, Xiaolong Wu, Benjamin Metcalfe, and Dingguo Zhang. "Effects of Different Preprocessing Pipelines on Motor Imagery-Based Brain-Computer Interfaces". In: *IEEE Journal of Biomedical and Health Informatics* 29.5 (May 2025), pp. 3343–3355. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2025.3532771.

[79] Marjan Urekar and Platon Sovilj. "EEG Dynamic Noise Floor Measurement with Stochastic Flash A/D Converter". In: *Biomedical Signal Processing and Control* 38 (Sept. 2017), pp. 337–345. ISSN: 17468094. DOI: 10.1016/j.bspc.2017.07.006.

[80] Jürgen Fell, Alexander Kaplan, Boris Darkhovsky, and Joachim Röschke. "EEG Analysis with Nonlinear Deterministic and Stochastic Methods: A Combined Strategy". In: *Acta Neurobiologiae Experimentalis* 60.1 (Mar. 31, 2000), pp. 87–108. ISSN: 1689-0035, 0065-1400. DOI: 10.55782/ane-2000-1328.

[81] Stefan Jonas, Michael Müller, Andrea O. Rossetti, et al. "Diagnostic and Prognostic EEG Analysis of Critically Ill Patients: A Deep Learning Study". In: *NeuroImage: Clinical* 36 (2022), p. 103167. ISSN: 22131582. DOI: 10.1016/j.nicl.2022.103167.

[82] Stephen M. Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E. Nichols, and Karla L. Miller. "Estimation of Brain Age Delta from Brain Imaging". In: *NeuroImage* 200 (Oct. 15, 2019), pp. 528–539. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2019.06.017.

[83] Wei-Long Zheng, Edilberto Amorim, Jin Jing, et al. "Predicting Neurological Outcome in Comatose Patients after Cardiac Arrest with Multiscale Deep Neural Networks". In: *Resuscitation* 169 (Dec. 2021), pp. 86–94. ISSN: 03009572. DOI: 10.1016/j.resuscitation.2021.10.034.

[84] Sérgio Baldo-Júnior, Murillo G. Carneiro, João L. M. Barbosa, et al. "A Bioinspired Multimodal CNN-LSTM Network for EEG Analysis of Patients in Coma". In: *Sensors* 25.22 (Jan. 2025), p. 6981. ISSN: 1424-8220. DOI: 10.3390/s25226981.

[85] Kartik K. Iyer, Sally Staton, Andrew Collaro, et al. "Decoding Brain Age Predictions from Sleep Electroencephalography across Infancy to Adolescence". In: *Scientific Reports* 15.1 (Nov. 24, 2025), p. 41590. ISSN: 2045-2322. DOI: 10.1038/s41598-025-25482-7.

[86] Edilberto Amorim, Michelle van der Stoel, Sunil B. Nagaraj, et al. "Quantitative EEG Reactivity and Machine Learning for Prognostication in Hypoxic-Ischemic Brain Injury". In: *Clinical Neurophysiology* 130.10 (Oct. 1, 2019), pp. 1908–1916. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2019.07.014.

[87] Velu Prabhakar Kumaravel, Elisabetta Farella, Eugenio Parise, and Marco Buiatti. "NEAR: An Artifact Removal Pipeline for Human Newborn EEG Data". In: *Developmental Cognitive Neuroscience* 54 (Apr. 1, 2022), p. 101068. ISSN: 1878-9293. DOI: 10.1016/j.dcn.2022.101068.

[88] Soheil Keshmiri. "Entropy and the Brain: An Overview". In: *Entropy* 22.9 (9 Sept. 2020), p. 917. ISSN: 1099-4300. DOI: 10.3390/e22090917.

[89] Lawrence J. Hirsch, Michael W. K. Fong, Markus Leitinger, et al. "American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2021 Version". In: *Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society* 38.1 (Jan. 1, 2021), pp. 1–29. ISSN: 1537-1603. DOI: 10.1097/WNP.0000000000000806. PMID: 33475321.

[90] *Pyragify: A Tool for Processing Code Repositories into Semantic Chunks for Analysis with LLMs, especiallyNotebookLM.* Version 0.2.0.

[91] *PleasePrompto/Notebooklm-Mcp: MCP Server for NotebookLM - Let Your AI Agents (Claude Code, Codex) Research Documentation Directly with Grounded, Citation-Backed Answers from Gemini. Persistent Auth, Library Management, Cross-Client Sharing. Zero Hallucinations, Just Your Knowledge Base.* URL: https://github.com/PleasePrompto/notebooklm-mcp (visited on 04/21/2026).

[92] Alois Schlögl and Gernot Supp. "Analyzing Event-Related EEG Data with Multivariate Autoregressive Parameters". In: *Progress in Brain Research*. Vol. 159. Elsevier, 2006, pp. 135–147. ISBN: 978-0-444-52183-5. DOI: 10.1016/S0079-6123(06)59009-0.

[93] Davoud Hajhassani, Quentin Barthélemy, Jérémie Mattout, and Marco Congedo. *Improved Riemannian Potato Field: An Automatic Artifact Rejection Method for Eeg*. 2025. DOI: 10.2139/ssrn.5165568. Pre-published.

# Glossary

**10-20 system** The international standard for electrode placement on the scalp, ensuring reproducible spatial coverage for EEG recordings by using distances of 10 percent and 20 percent between anatomical landmarks.

**Analog-to-Digital Converter (ADC)** The hardware component that translates continuous scalp-voltage fluctuations into discrete numerical values for digital EEG processing.

**aperiodic activity** The non-oscillatory $1/f$-like spectral component of the EEG, characterised by its slope (exponent) and offset; changes in aperiodic activity confound traditional band-power measures.

**best CPC** The best (lowest) Cerebral Performance Category score recorded for a patient within the follow-up period, typically 6 months post-arrest, used to mitigate transient fluctuations in early assessments.

**biological ghost** A phenomenon identified in this thesis where fundamental biometric traits, such as biological sex, remain detectable only as residual traces ($AUC \approx 0.57$) following the near-total erasure of biological variance by hypoxic-ischaemic injury.

**Brain Age Gap (BAG)** The numerical difference between a subject's chronological age and the age predicted by an EEG-based machine learning model. It serves as a proxy for accelerated biological ageing or cumulative neural frailty.

**burst-suppression** An EEG pattern characterised by alternating epochs of high-amplitude activity (bursts) and near-isoelectric quiescence (suppressions), commonly observed after severe cerebral hypoxia.

**cardiac contamination** The presence of ECG-derived signals (QRS complexes, pulse artefacts) in EEG recordings; in the ICU context, these can dominate low-amplitude comatose EEG and introduce confounding spectral peaks.

**Cerebral Performance Category (CPC)** Ordinal neurological outcome scale ranging from 1 (good cerebral performance) through 3 (severe cerebral disability) to 5 (death); used as the primary outcome measure in post-cardiac-arrest prognostication research.

**Coma-Normative Brain Age Gap** A decision-independent biomarker variant developed in this thesis. It measures an individual's deviation from the age-expected EEG patterns within the comatose population, circumventing the rejuvenation artefact of healthy-norm models.

**confounder inseparability** The condition under which permanent cortical damage, reversible pharmacological and physiological state effects (sedation, targeted temperature management), and premorbid neurobiological traits (including chronological age) produce overlapping EEG signatures that cannot be disentangled from the recorded signal alone.

**constant-predictor signature** The pattern where a regression model outputs approximately the same value, near the training-set mean, for all inputs, yielding an R-squared of approximately zero.

**Global Interpreter Lock (GIL)** A mutex in the CPython interpreter that prevents multiple native threads from executing Python bytecodes at once, necessitating process-based parallelism for CPU-bound EEG tasks.

**High-Performance Computing (HPC)** The use of parallel processing and cluster architectures to execute computationally intensive tasks, such as large-scale EEG feature extraction or deep learning training.

**Hypoxic-Ischaemic Encephalopathy (HIE)** Brain injury resulting from a period of insufficient oxygen (hypoxia) and blood flow (ischaemia), typically following cardiac arrest; severity ranges from mild cortical dysfunction to complete neuronal death.

**Hypoxic-Ischaemic Encephalopathy (HIE)** A brain injury caused by a systemic lack of oxygen and blood flow, commonly occurring after cardiac arrest, leading to a spectrum of neurological outcomes from full recovery to brain death.

**In-hospital cardiac arrest (IHCA)** Cardiac arrest occurring within the hospital, typically characterised by shorter downtime, immediate initiation of advanced life

support, and a higher prevalence of metabolic derangements compared to community arrests.

**normative modelling** A statistical framework that establishes a reference distribution from a normative (typically healthy) cohort and quantifies how far an individual deviates from that reference.

**Out-of-hospital cardiac arrest (OHCA)** Cardiac arrest occurring in the community, often associated with longer durations of untreated hypoxia and a specific ischaemic-reperfusion injury profile that differs from IHCA trajectories.

**preprocessing profile** A declarative configuration of signal-processing transformations and quality gates applied to the raw multi-channel signal.

**quality mask** A non-destructive bitfield array used to annotate hardware faults and physiological artefacts without altering the underlying signal dimensions.

**rejuvenation artefact** The observation that burst-suppression and other severe encephalopathy patterns produce EEG spectra that healthy-brain-age models misinterpret as *young*, systematically reversing the expected direction of the brain age gap.

**Return of Spontaneous Circulation (ROSC)** The resumption of sustained heart rhythm and blood flow following a period of cardiac arrest, usually marked by the presence of a palpable pulse or measurable blood pressure.

**self-fulfilling prophecy** In neuroprognostication, a circular bias where a prediction of a poor outcome leads to the withdrawal of life-sustaining therapy. This clinical action ensures the predicted death occurs, creating an artificial true positive that inflates model performance metrics.

**state-trait eclipse** The physiological phenomenon in which acute pathological states, such as profound sedation or severe hypoxic-ischaemic injury, fundamentally obscure the stable neurophysiological traits—including biological age and sex—that would otherwise be detectable in healthy EEG.

**Withdrawal of Life-Sustaining Treatment (WLST)** The clinical decision to cease interventions such as mechanical ventilation or vasopressor support when the prognosis is deemed futile, often the primary cause of death in post-cardiac arrest cohorts.

# Acronyms

**AED** Automated External Defibrillator

**aEEG** Amplitude-integrated Electroencephalography

**ApEn** Approximate Entropy

**AUC** Area Under the ROC Curve

**BAEP** Brainstem Auditory Evoked Potentials

**BiLSTM** Bidirectional Long Short-Term Memory

**BSR** Burst-Suppression Ratio

**CA** Cardiac Arrest

**cEEG** Continuous Electroencephalography

**CI** Confidence Interval

**CNN** Convolutional Neural Network

**CPR** Cardiopulmonary Resuscitation

**CPU** Central Processing Unit

**CV** Cross-Validation

**DFA** Detrended Fluctuation Analysis

**ECG** Electrocardiography

**ECMO** Extracorporeal Membrane Oxygenation

**EECI** EEG ECG Contamination Index

**EEG** Electroencephalography

**ElasticNet** Elastic-Net Regularised Linear Regression

**EMG** Electromyography

**EMS** Emergency Medical Services

**EOG** Electrooculography

**FFT** Fast Fourier Transform

**FOOOF** Fitting Oscillations and One-Over-F

**FWER** Family-Wise Error Rate

**GBM** Gradient Boosting Machine

**GCS** Glasgow Coma Scale

**GIL** Global Interpreter Lock

**GOS** Glasgow Outcome Scale

**GPD** Generalised Periodic Discharges

**GPU** Graphics Processing Unit

**Grad-CAM** Gradient-weighted Class Activation Mapping

**GRDA** Generalised Rhythmic Delta Activity

**GRU** Gated Recurrent Unit

**HFO** High-Frequency Oscillations

**HPC** High-Performance Computing

**I-CARE** International Cardiac Arrest REsearch Consortium

**ICA** Independent Component Analysis

**ICC** Intraclass Correlation Coefficient

**ICU** Intensive Care Unit

**IHCA** In-hospital cardiac arrest

**IQR** Interquartile Range

**LightGBM** Light Gradient-Boosting Machine

**LOPO** Leave-One-Patient-Out Cross-validation

**LSTM** Long Short-Term Memory

**MAE** Mean Absolute Error

**MLOps** Machine Learning Operations

**MLP** Multi-Layer Perceptron

**mRS** Modified Rankin Scale

**MSE** Mean Squared Error

**NSE** Neuron-Specific Enolase

**OHCA** Out-of-hospital cardiac arrest

**PCA** Principal Component Analysis

**PLV** Phase Locking Value

**PSD** Power Spectral Density

**qEEG** Quantitative Electroencephalography

**RANSAC** RANdom SAmple Consensus

**RF** Random Forest

**RMSE** Root Mean Squared Error

**ROC** Receiver Operating Characteristic

**ROSC** Return of Spontaneous Circulation

**RTM** Regression to the Mean

**SampEn** Sample Entropy

**SD** Standard Deviation

**SE** Standard Error

**SEF** Spectral Edge Frequency

**SHAP** SHapley Additive exPlanations

**SLURM** Simple Linux Utility for Resource Management

**SSEP** Somatosensory Evoked Potentials

**SVM** Support Vector Machine

**TCN** Temporal Convolutional Network

**TTM** Targeted Temperature Management

**VRAM** Video RAM

**WFDB** WaveForm DataBase

**WLST** Withdrawal of Life-Sustaining Treatment

**wPLI** Weighted Phase Lag Index

**XGBoost** Extreme Gradient Boosting

# Guide to the Appendices

The five appendices collect all technical reference material, keeping the main narrative uncluttered. Each appendix is self-contained and can be read independently; the table below gives a one-line orientation.

| App. | Title | Contents |
|---|---|---|
| A | Signal Processing & Feature Engineering | Preprocessing profiles, two-stage quality control, empirical threshold validation, full feature catalogue, hierarchical aggregation, domain-cluster analysis, and feature EDA. |
| B | Experimental Framework & Full Results | **Part A**: RO→hypothesis mapping, experiment dependency graph, summary table. **Part B**: pipeline implementation, run configurations, hyperparameter grids, Phase 2 architectures. **Part C**: Phase 1/2 results, CV diagnostics, H-1–H-11 evaluation, model-collapse diagnostics. |
| C | Project Governance, Reproducibility & Infrastructure | Module taxonomy, formal process definitions, HPC and storage infrastructure, testing & QA, monitoring dashboards, AI-workflow and citation-gate rules. |
| D | Formal Bayesian Proof | Derivation of the self-fulfilling-prophecy bias coefficient underlying the decision-independent framing of CN-BAG. |
| E | Statistical Analysis Plan | Pre-registered hypothesis register, multiple-comparison strategy, power analysis, and severity-score validation framework. |

# Appendix A

# Signal Processing and Feature Engineering

This appendix provides the technical reference for the signal-processing and feature-engineering stack, following the data-flow from raw waveform to model-ready representation. Section A.1 documents the five preprocessing profiles and parallel data-routing architecture. Section A.2 covers the two-stage quality-control system. Section A.3 presents empirical threshold validation. Sections A.4 and A.5 document the full feature catalogue and the feature-adaptive hierarchical aggregation scheme. Section A.6 analyses domain clusters. Section A.7 provides extended feature-distribution analyses.

## A.1 Preprocessing Profiles and Data Routing

The extraction and quantification of neurophysiological traits in comatose patients necessitate strict control over data modification. This methodology routes the continuous electroencephalogram through discrete, parallel preprocessing profiles to systematically isolate the effects of algorithmic cleaning. The implementation architecture supporting this parallel routing is shown in Figure A.1.

### A.1.1 Profile Specifications (A–E)

Table A.1 provides the complete specification of the five preprocessing profiles (A–E and the C-noecg ablation variant) applied across the cohort. Figure A.2 provides the visual side-by-side comparison of the signal output from each profile; the compact repository-level CLI and execution reference is provided in Appendix B (Section B.13.1).

**Table A.1:** Preprocessing profile specifications (A–E and C-noecg ablation variant). All profiles resample to 256 Hz. Filter type is IIR Butterworth 4th order (zero-phase via `sosfiltfilt`) unless noted.

| Profile | HPF | LPF | Notch | Reference | ICA | Notes |
|---|---|---|---|---|---|---|
| A | — | — | — | CAR | — | Raw/resampled baseline (RO-3). No filtering, no artefact removal. Quality mask tracks corrupt samples. |
| B | 0.5 Hz | 45 Hz | Adaptive | Bipolar | — | Phase 1 ML profile. Step removal, bad-channel interpolation. Signal in $\mu$V (no normalisation). Features on referential + bipolar. |
| C | 0.1 Hz | — | Adaptive | CAR | — | Phase 2 DL baseline. ECG *retained* by design (no ICA). Robust normalisation ($\pm 10$ clip). DL-safe NaN fill. |
| C-noecg | 0.1 Hz | — | Adaptive | CAR | $\lvert r\rvert > 0.90$ | Matched ablation: identical to C plus ECG-ICA + OBS fallback. Output is ECG-removed signal. |
| D | PREP pipeline | | PREP | Bipolar | ICLabel | PREP bad-channel detection (RANSAC disabled for coma EEG) + ICLabel classification. Robust normalisation. |
| E | 0.5 Hz | 45 Hz | — | CAR (all) | — | MNE textbook defaults. No bad-channel detection, no ICA. Robust normalisation. |


**Figure A.1:** Repository implementation screenshot for the preprocessing profile pipeline runner. The figure anchors the implementation context discussed in this section: profile selection (A, B, C, C-noecg, D, E), patient-scoped execution, and reproducible CLI-driven orchestration from a single module entry point.

## A.1.2 Filter Specifications

### High-Pass Filter

- Type: IIR Butterworth, 4th order
- Cutoff: Profile B/E = 0.5 Hz; Profile C/C-noecg = 0.1 Hz
- Phase: Zero-phase (`sosfiltfilt`, forward–backward pass)
- NaN safety: corrupt samples interpolated before filtering, restored afterward (preserving quality mask)
- Mask expansion after filtering to exclude IIR ringing at valid/invalid boundaries: Profile B/E ($f_c$=0.5 Hz) $\approx$ 1.3 s/edge; Profile C ($f_c$=0.1 Hz) $\approx$ 6.4 s/edge
- Baseline correction is limited to the HPF; slow drift is annotated in Stage 2 rather than subtracted.

### Low-Pass Filter

- Type: IIR Butterworth, 4th order

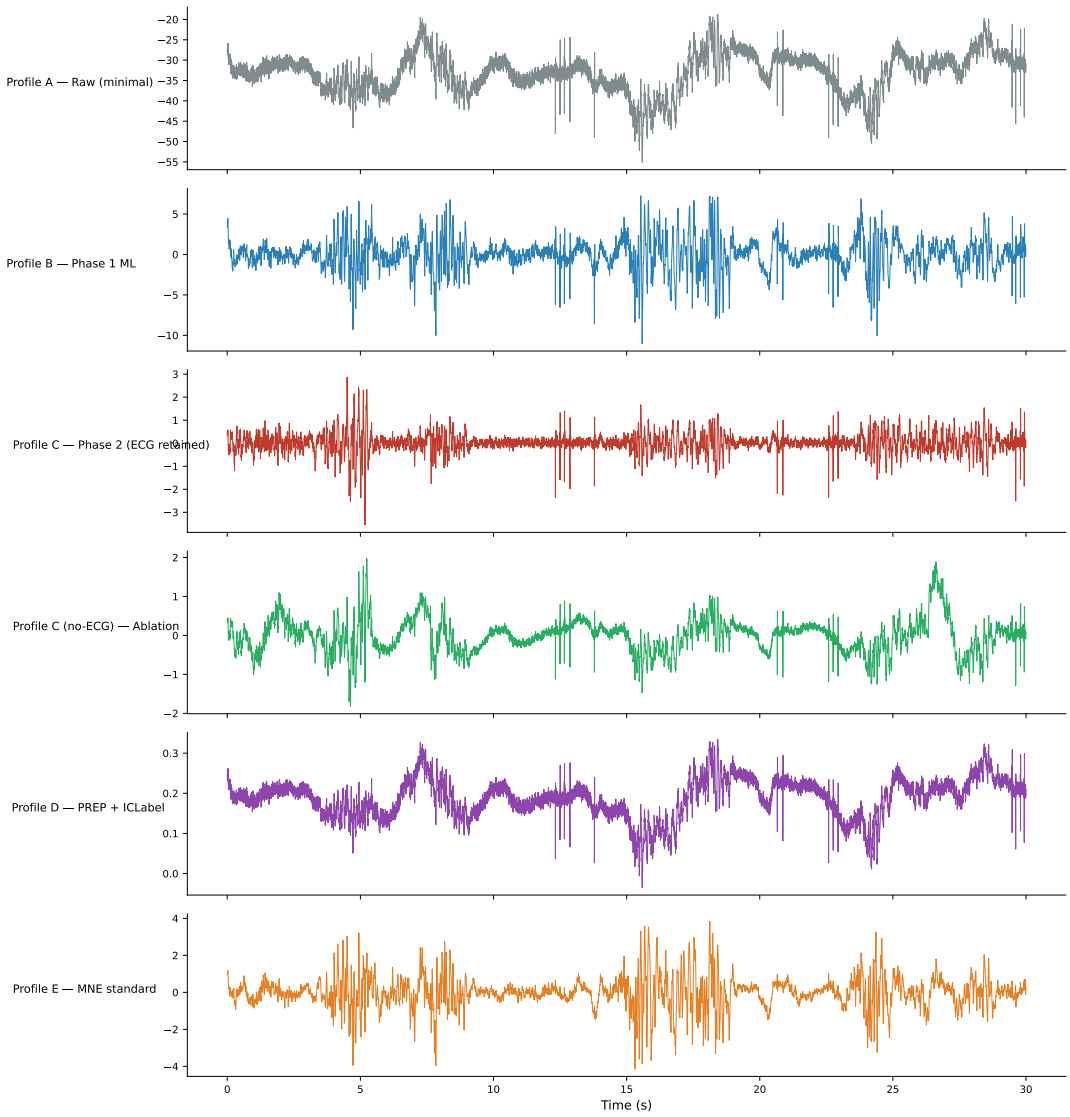



**Figure A.2:** Side-by-side comparison of the five preprocessing profiles (A–E) applied to the same recording segment. Figure A.2 illustrates the progressive cleaning from raw (Profile A) through the retained-ECG Phase 2 baseline (Profile C), the matched ECG-removal ablation (Profile C-noecg), and PREP-based processing (Profile D). In Profile C, cardiac artefact is clearly visible as regular high-amplitude spikes because no ICA ECG removal is applied at this stage.

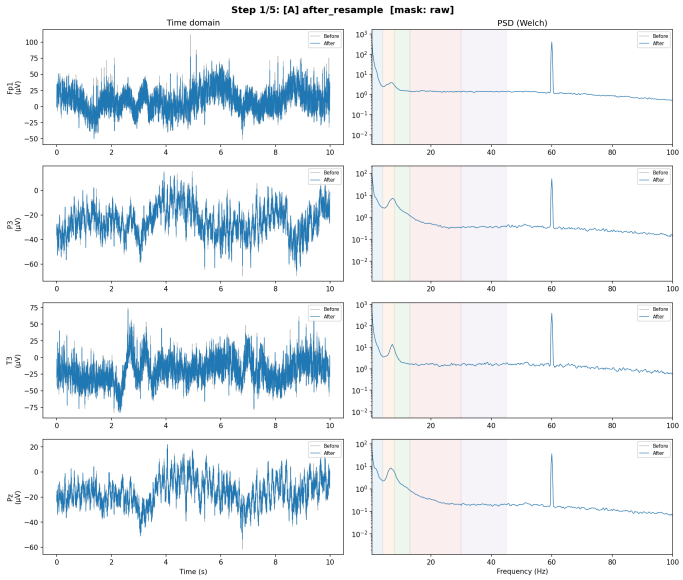
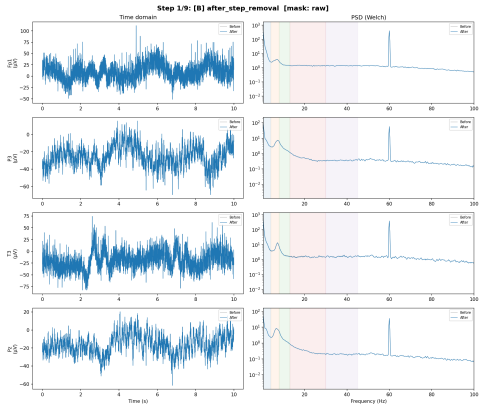



**Figure A.3:** Profile A applies only anti-aliasing resampling to the raw 10-20 EEG signal, retaining all artefacts. This forms the empirical baseline for threshold validation (`EXP-001-A`), as artefact distributions must be profiled in the unprocessed state before rejection thresholds can be justified. Bipolar double-banana montage, patient 0286.

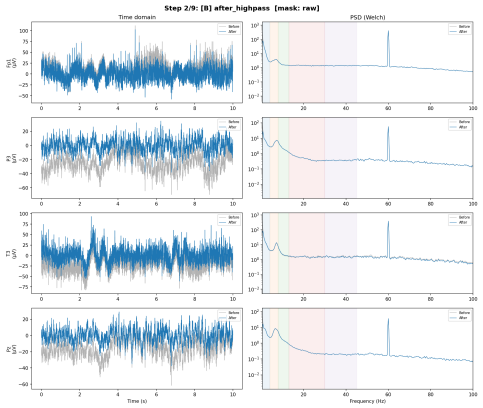
- Cutoff: 45 Hz (Profiles B and E)
- Phase: Zero-phase (`sosfiltfilt`)

### Notch Filter

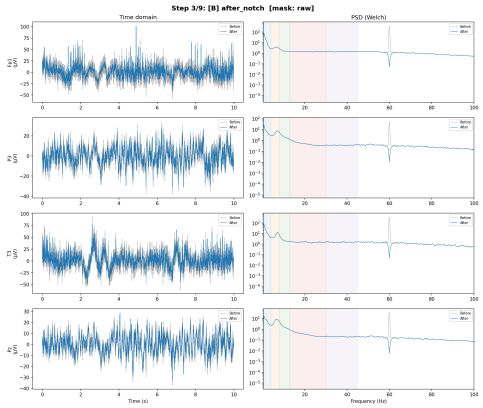
- **Frequency:** Taken from the WFDB header `utility_freq` field (50 or 60 Hz), not auto-detected from the PSD. Auto-detection was rejected because low-amplitude comatose EEG makes the line-noise vs. brain-signal ratio ambiguous. When the header field is absent, the frequency defaults to 50 Hz.
- **Prominence gate (apply vs. bypass):** A Welch PSD is computed from quality-masked samples (only `mask == 0x00` samples contribute). When the spectral peak prominence at the line frequency falls below 3 dB (i.e. line noise is indistinguishable from the spectral floor), the notch is bypassed entirely.
- **Q-factor estimation:** $Q = f_{\text{line}}/\Delta f_{3\text{dB}}$, clamped to $[20, 80]$. The estimator measures the half-power width of the PSD peak and maps it to a continuous Q value. However, with 4-second Welch windows the frequency resolution is $\approx 0.25$ Hz, and mains-frequency peaks are typically narrower than one bin. When



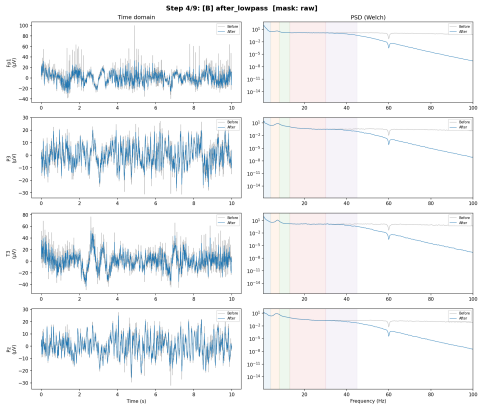

**(a)** After step-artefact removal



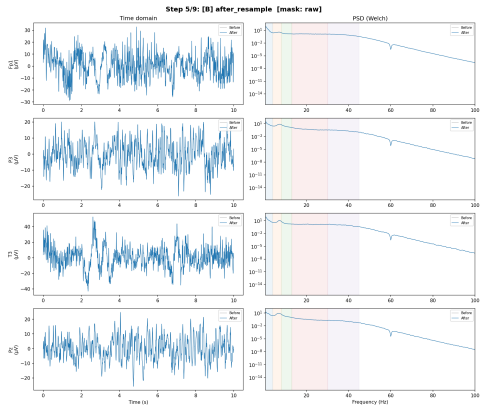

**(b)** After 0.5 Hz high-pass filter




**(c)** After 50/60 Hz IIR notch




**(d)** After 45 Hz low-pass filter




**(e)** After resample to 256 Hz (final output)

**Figure A.4:** Profile B stepwise signal transformation for patient 0286. The bipolar double-banana montage is retained throughout all stages (no re-referencing is applied). The 0.5 Hz high-pass filter removes slow drift; the IIR notch suppresses mains interference; the 45 Hz low-pass removes high-frequency muscle artefact; the final resampling standardises the sample rate to 256 Hz. Profile B outputs are used directly without normalisation for Phase 1 classical modelling, where absolute amplitude magnitude carries neurophysiological meaning (in contrast to Profile C, which applies temporal CAR and patient-level robust normalisation; see Figure 4.6 in the main text).

the true peak width falls below the frequency-resolution floor (0.25 Hz), this yields $Q = 50/0.25 = 200$ (or $60/0.25 = 240$), which clips to $Q_{\max} = 80$. The Q adaptation is therefore implemented but resolution-limited for narrow peaks; in practice the upper clamp is frequently reached in the I-CARE cohort (see Figure A.5). The choice of $Q_{\max} = 80$ is a design parameter that was not systematically optimised and represents a potential area for improvement (see Section 6.7).

- **Filter type:** Second-order IIR Butterworth notch, selected at design time by EXP-306 (463/483 patient-level votes favouring IIR over FIR).
- **Harmonics:** The filter is applied at up to 3 harmonics ($f$, $2f$, $3f$), provided each harmonic remains below the Nyquist frequency.

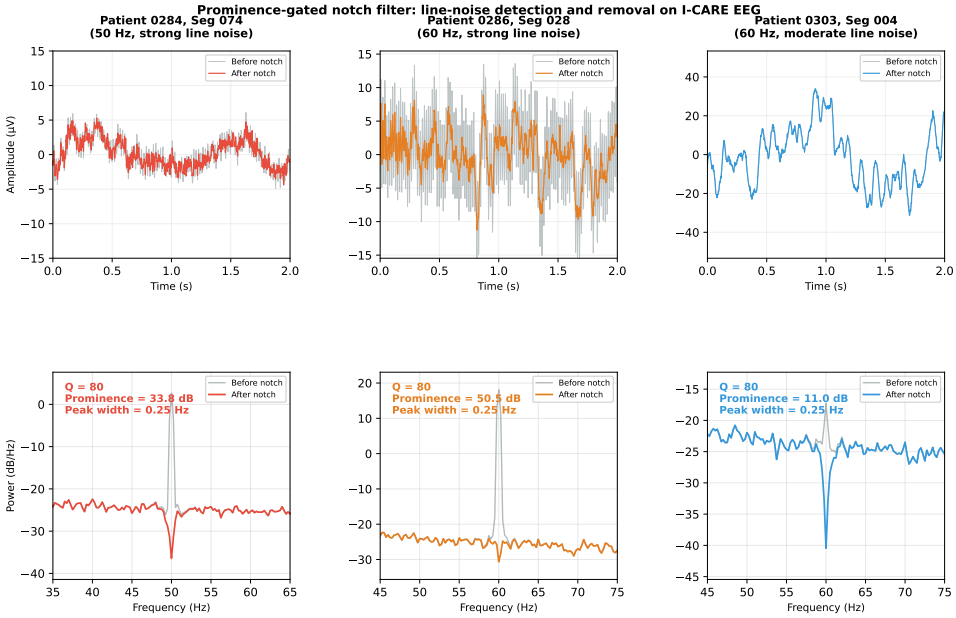



**Figure A.5:** Prominence-gated notch filter on three I-CARE patient segments with different line-noise severity and line frequencies, each preceded by a 0.5 Hz Butterworth high-pass as in the production pipeline. Left: strong 50 Hz peak (prominence 34 dB). Centre: strong 60 Hz peak (prominence 51 dB). Right: moderate 60 Hz peak (prominence 11 dB). In all three cases the notch fires at $Q = 80$, illustrating that the Q-factor saturates at its upper clamp because the Welch frequency resolution ($\approx 0.25$ Hz) cannot resolve the intrinsic peak width. The functional adaptive element is the prominence gate: segments below 3 dB are left unfiltered. Top row: 2 s time-domain overlay. Bottom row: Welch PSD zoomed to $\pm 15$ Hz around the line frequency.

### Step Removal (DC Offset Repair)

- Threshold: per-sample jump $> 50$ $\mu$V/sample flags a step candidate.

- Estimation: DC offset is computed from short windows before/after the step (default 50 samples each side), then removed via cumulative correction.
- Quality-mask guard: if $> 50\%$ of the estimation window is masked, the step is skipped to avoid misclassifying saturation–zero transitions.
- Logging: per-channel counts, locations, and offset sizes are recorded for QA.

### A.1.3 Bipolar Montage Specification

The bipolar (double-banana) montage used in Profiles B and E comprises 10 channel pairs following standard clinical practice:

| Chain | Pairs |
|---|---|
| Left temporal | Fp1–F7, F7–T3, T3–T5, T5–O1 |
| Right temporal | Fp2–F8, F8–T4, T4–T6, T6–O2 |
| Central | Fz–Cz, Cz–Pz |

### A.1.4 Pipeline Step Ordering Rationale

The ordering of preprocessing steps within each profile is not arbitrary; each sequencing decision reflects a specific domain constraint. The key choices are documented below.

**Quality mask before all filtering.** The raw quality mask is computed from the unmodified signal before any filter is applied. IIR filters spread transient artefacts (electrode pops, saturation cliffs) into adjacent samples; computing the mask post-filter would underestimate true data loss and misattribute filter ringing to physiology. The mask is therefore frozen before filtering and only expanded post-filter to annotate IIR settling zones (Section A.1.2), not to add new fault detections.

**Linear interpolation before IIR filtering.** Corrupt segments flagged by the quality mask are linearly interpolated before the IIR filter processes the signal. This prevents masked zeros or transient spikes from generating filter ringing that propagates into clean adjacent regions. After filtering, the interpolated values are discarded and the mask is restored; the output retains `NaN` at all originally masked positions.

**HPF before notch (Profiles B, C, and C-noecg).** The high-pass filter is applied before the notch in all profiles that include both steps. A large DC offset raises the low-frequency PSD floor and biases the prominence estimate used by the notch filter's bypass gate. Removing DC and slow drift first produces a cleaner power spectrum for the detection decision.

**Profile B: notch followed by low-pass.** Profile B applies the prominence-gated notch filter at the WFDB-header line frequency before the 45 Hz low-pass. Although the low-pass alone provides $> 40$ dB attenuation at 50 and 60 Hz, the notch is retained because the LPF transition band (45–60 Hz) would otherwise carry residual line-noise energy into the resampled signal. The combined notch + LPF ensures a clean spectral floor for the Phase 1 ML feature extractors.

**Profile C: no low-pass filter.** Phase 2 deep learning models learn representations end-to-end and benefit from retaining full spectral content up to the Nyquist limit. Anti-aliasing at 256 Hz is handled internally by `resample_poly`'s Kaiser-windowed FIR. An explicit 45 Hz low-pass would truncate potentially informative high-frequency content before the DL feature extractor can evaluate it.

**Temporary CAR before ECG-ICA (Profile C-noecg).** ICA decomposition assumes a zero-mean reference. In referential recordings, the cardiac artefact appears as a distributed but uncentred signal. A temporary common average reference (CAR), computed only over valid (unmasked) channels, centres the signal for ICA. The temporary CAR is discarded after ICA decomposition; the pipeline output retains the original referential montage corrected only for the extracted ECG component. Profile C does not apply ICA; this step is specific to Profile C-noecg.

**RANSAC bad-channel detection disabled for coma EEG (Profile D).** RANSAC predicts each channel from its spatially weighted neighbours and flags channels with high prediction error; this mechanism assumes that healthy channels are mutually predictable (typically $r > 0.75$ in resting adult EEG). In comatose EEG, pathological cortical disconnection reduces inter-channel correlation to moderate and variable levels (median $r \approx 0.46$, IQR 0.37–0.58; EXP-306), reflecting genuine loss of neural connectivity rather than electrode fault. RANSAC's spatial prediction degrades at these correlation levels and cannot reliably separate defective from pathological channels; RANSAC was there-

fore disabled (`do_ransac=False`) for Profile D. Additionally, the PREP correlation criterion (flagging channels with $r < 0.40$ in $>1\,\%$ of windows) fires on pathological-but-functioning channels whose connectivity has been reduced by brain injury, systematically confusing neurological damage with electrode failure. Bad-channel detection for Profile D therefore falls back to the absolute amplitude and variance thresholds of the Stage 1 quality mask.

**Profile E: unmasked CAR by design.** Profile E implements MNE textbook defaults without domain adaptation. The unmasked CAR averages across all 19 channels regardless of quality status. This is a deliberate naive baseline demonstrating the artefact propagation that results when quality-mask awareness is omitted, thereby providing the RO-3 quantitative benchmark against domain-adapted alternatives.

### A.1.5 ICA Configuration

ICA decomposition is applied only in Profiles C-noecg (ECG-correlation ICA) and D (ICLabel ICA). The following subsections detail the shared decomposition settings and the profile-specific configuration.

#### ICA Decomposition (Shared Settings)

- Algorithm: Extended InfoMax via Picard solver (`method='picard', fit_params=dict(ortho=False, extended=True)`)
- Rationale: Picard handles mixed sub-/super-Gaussian EEG source distributions and matches ICLabel's training domain
- Temporary 1 Hz high-pass applied for decomposition stability; unmixing matrix applied to the profile's lower-cutoff signal
- Random state: 42

#### Profile C-noecg: ECG-Correlation ICA

- Targets cardiac artefact by correlating each ICA component with the recorded ECG channel
- Correlation threshold: $|r| > 0.90$ (conservative; only confidently cardiac components removed)

- Fallback: template-based ECG detection via autocorrelation when no ECG channel is present
- If ICA removes 0 components (common in coma EEG where cardiac signal is not spatially separable), falls back to OBS template subtraction

**Profile D: ICLabel ICA**

- MNE-ICLabel neural network classifies components as brain / muscle / eye / heart / line_noise / channel_noise / other
- Conservative thresholds for comatose EEG: muscle/eye/heart = 0.70; line/channel noise = 0.75
- Maximum rejected components: 5
- Caveat: ICLabel trained on healthy awake EEG only; accuracy on burst suppression and periodic discharges is unvalidated

**Quality Checks (All ICA Methods)**

- Broadband power change $< 10\%$
- Brain band (1–30 Hz) preserved $\pm 10\%$
- Overall variance reduction $< 30\%$

## A.2 Two-Stage Quality Architecture

To ensure robust feature extraction without compromising the underlying neurophysiological state, the pipeline implements a strictly parameterised two-stage quality control architecture. Stage 1 enforces absolute biophysical limits to mask hardware faults, whilst Stage 2 annotates residual signal patterns without relying on destructive spatial interpolation.

### A.2.1 Stage 1: Hardware Fault Masking

The first stage operates exclusively on fundamental physical boundaries before any destructive resampling or filtering occurs. Any signal interval exceeding physiological plausibility is strictly masked using binary exclusion arrays.

**Table A.2:** Stage 1 quality mask thresholds (raw quality mask, sample-level). Code defaults from `compute_raw_quality_mask`; empirically validated via EXP-306.

| Defect | Threshold | Code default | Rationale |
|---|---|---|---|
| NaN / Inf | `!isfinite(x)` | — | Non-numeric ADC output |
| Exact zeros | $\|x\| < 10^{-15}$, $\geq 2$ consecutive | — | Equipment dropout |
| Flatline | std $< 0.01$ $\mu$V | 100 ms sliding window | ADC noise floor |
| Jump | $\|\Delta x\| >$ 200 $\mu$V/sample | — | Normal EEG $< 50$ $\mu$V/sample |
| Electrode pop | peak $> 150$ $\mu$V (median-subtracted) | $\pm$20 ms dilation | Fills gap between jump and saturation |
| Saturation | $\|x\| > 500$ $\mu$V | — | Amplifier clipping |

### A.2.2 NaN-Guard and Quality Mask Schema

To preserve structural missingness as an explicit signal the extractor implements a small, well-defined "NaN-guard" wrapper around every numeric operation. The guard never replaces missingness with silent defaults; instead it propagates a `quality_mask` bitfield alongside numeric arrays and emits per-window `nan_fraction` metadata used downstream by both Phase 1 and Phase 2 models.

```
def nan_guarded_apply(signal, op, quality_mask):
        # signal: numpy array (samples x channels)
        # op: function to apply to valid segments (e.g., filter, PSD)
        # quality_mask: uint8 bitfield per sample (same length as
    signal)

        valid = is_finite(signal) & (quality_mask == 0)
        segments = contiguous_true_segments(valid)
        out = empty_like(signal) * np.nan
        for (start, end) in segments:
                window = signal[start:end]
                out[start:end] = op(window)
        # Preserve quality mask and compute nan_fraction per-window
        nan_fraction = compute_nan_fraction_per_window(out)
        return out, quality_mask, nan_fraction
```

**Listing A.1:** Pseudocode: NaN-Guard wrapper (conceptual)

The `quality_mask` is stored as a `uint8` bitfield per-sample. A canonical mapping used throughout the codebase is shown in Table A.3; the associated Stage 1 decision rules

are detailed in Section A.2.

**Table A.3:** Quality mask bitfield mapping (`uint8`). Bit positions are constant across profiles and persisted in Parquet metadata.

| Bit (LSB=0) | Name | Meaning |
|---|---|---|
| 0 | NAN_INF | Non-finite sample (NaN or Inf) |
| 1 | EXACT_ZERO | Exact zeros (multiple consecutive) |
| 2 | FLATLINE | Low-variance flatline segment |
| 3 | JUMP | Large per-sample jump (>200 µV) |
| 4 | SATURATION | Amplitude > 500 µV |
| 5 | ELECTRODE_POP | Electrode pop / transient |
| 6 | RESERVED | Reserved for future use |
| 7 | RESERVED | Reserved for future use |

An example Parquet metadata row from the feature extraction output (abridged) is:


```
{
    "patient_id": "0284",
    "profile": "C",
    "segment_id": "0284_0001",
    "quality_mask_bits": "uint8[]",
    "nan_fraction": 0.012,
    "git_hash": "823106f3d"
}
```


### A.2.3 Stage 2: Post-Filter Artefact and Pattern Detection

Stage 2 operates on the filtered signal and produces structured descriptors that do not modify the quality mask.

Where Stage 1 uses a `uint8` bitfield to record hardware faults, Stage 2 employs a `uint16` bitmask array to annotate physiological artefacts (e.g. EMG, ECG) and neurological patterns (e.g. burst-suppression, seizures) post-filtering. The expanded 16-bit capacity allows a wide array of independent clinical and technical states to be encoded simultaneously via bitwise OR. Crucially, these patterns are annotated as metadata rather than deleted: downstream analysis dynamically layers the Stage 2 mask on top of the Stage 1 quality mask to apply study-specific exclusions (e.g. excluding burst-suppression for alpha-peak analysis but retaining it for brain-age prediction). This design means

**Table A.4:** Stage 2 annotation taxonomy and operational definitions. These Stage 2 patterns are flagged as metadata but never used to delete the underlying signal, following the 'quantify-not-remove' philosophy of the thesis.

| Category | Pattern Name | Operational Definition |
|---|---|---|
| Residual Artefacts | EMG (muscle) | broadband power $> 20$ Hz |
| | Residual ECG | cardiac-rate periodic component |
| | Electrode pop | post-filter jumps |
| | HF noise | $> 45$ Hz z-score |
| | EOG | frontal channels, $> 100$ $\mu$V |
| | Baseline drift | slow trend, $z > 2$ |
| | Saturation | $\lvert x\rvert > 500$ $\mu$V after processing |
| Neurological Signals | Burst suppression | alternating burst/suppression periods |
| | Low amplitude | continuous low amplitude (no bursts) |
| | Spindle coma | 11–16 Hz sigma Hilbert plus fusiform morphology |

researchers can slice the dataset by pathology or artefact type without permanently destroying the continuous waveform arrays.

### Per-Window Bad Channel Detection

Bad-channel detection operates at the per-window level and flags individual electrode segments rather than removing entire channels from the recording. Applied after Stage 1 masking; windows with $< 50\%$ valid samples are skipped.

- Flatline: std $<$ threshold **AND** MAD $<$ threshold (conservative double criterion for comatose EEG)
- Saturation: $\max \lvert x\rvert >$ sat_threshold
- Excessive zeros: $> 50\%$ zero-valued samples
- Line noise: power ratio at utility frequency $\pm 2$ Hz exceeds threshold
- Inter-channel correlation: computed and stored for analysis but *not used for channel flagging* — comatose EEG has inherently low inter-channel correlation (median ICC $= 0.46$, EXP-306) that is expected and not indicative of bad channels

## A.3 Empirical Threshold Validation

The deployment of absolute physical bounds necessitates rigorous empirical validation. This section documents the empirical grounding of the rejection criteria against development-cohort distributions.

### A.3.1 Threshold-Specific Validation

- **Flatline threshold (0.01 $\mu$V):** CDF analysis of the development cohort confirmed a clear distributional gap between ADC noise-floor samples and physiological low-amplitude coma EEG. The threshold sits within this gap: the 5th-percentile IQR of genuine low-amplitude signals is approximately 2.8 $\mu$V, some 280$\times$ above the 0.01 $\mu$V limit.
- **Saturation threshold ($> 500$ $\mu$V):** CDF analysis confirmed that 500 $\mu$V corresponds to the 75th percentile of per-patient amplitude maxima across the development cohort; values above this level represent hardware saturation rather than genuine neural signal.
- **Jump threshold ($> 200$ $\mu$V/sample):** This threshold is grounded in a hard amplifier slew-rate ceiling rather than a distributional gap. The inter-sample jump distribution is continuous (90th percentile $\approx 155$ $\mu$V/sample; 95th percentile $\approx 317$ $\mu$V/sample); no CDF gap was observed. The bound excludes only samples that are physically impossible given the recording hardware.
- **Dead-channel criterion:** Validated via hardware argument; the probability of true biological zeros is $< 10^{-9}$.

### A.3.2 Dataset Quality Characterisation (EXP-012)

Profile A preprocessing (raw/resampled, no filtering) was applied to the full development set ($N = 483$ patients) to characterise the technical quality of the I-CARE EEG recordings before any signal processing.

**Table A.5:** I-CARE EEG recording quality characterisation from EXP-012 ($N = 483$, Profile A). Values represent the proportion of patients with at least one affected recording segment.

| Artefact / Quality Issue | Prevalence (%) | Notes |
|---|---|---|
| DC offset (baseline drift) | 62.7 | Electrochemical polarisation / electrode movement |
| Step artefact damage | 60.5 | DC step transients causing filter ringing |

## A.4 Comprehensive Feature Catalogue

This section catalogues the metrics extracted to characterise the cortical dynamics. The full extraction pipeline produces 3,423 base features across eight extractor modules when applied to Profile B/D (monopolar and bipolar channels combined; Table A.7). Profile A/C/E (monopolar channels only, 19 referential) extracts 1,578 features; the 1,845-feature increment in Profile B/D arises from computing spectral, aperiodic, and temporal features across ten additional bipolar channel pairs prior to post-extraction deduplication. The ten primary age-relevant feature families are listed in Table A.6 with their scientific rationale and ACNS alignment; the complete per-module counts and the deduplication step are summarised in Table A.7.

**Table A.6:** CN-BAG pipeline feature extractor families (Tier 2, Phase 1). *Age relevance* indicates the primary scientific rationale for including the family in the brain-age regression; *ACNS alignment* notes the closest ACNS 2021 Critical Care EEG Terminology category.

| Family | Window | ∼N | Age relevance | ACNS alignment |
|---|---|---|---|---|
| `core_spectral` | 30 s | ∼540 | PAF (primary age marker); aperiodic exponent ($1/f$; age effect, ICU caveat); absolute and relative band powers | Background frequency, voltage |
| `core_temporal` | 5 s | ∼135 | Hjorth mobility (dominant frequency proxy); kurtosis | Continuity, amplitude |
| `core_nonlinear` | 60 s | ∼135 | SampEn, DFA, Higuchi FD — complexity declines with both age and injury | — |
| `burst` | 120 s | ∼114 | State covariate; BSR reflects coma depth, not age; useful as confound regressor | Suppression ratio, continuity class |
| `aeeg` | 120 s | ∼114 | Amplitude proxy; lower margin sensitive to suppression; upper margin to paroxysmal events | Background amplitude |
| `additional` | 60 s | ∼156 | BIS approximation (beta/sigma ratio); beta/gamma carry sex-effect confounders | — |
| `pathological_fn` | 60 s | ∼25 | State descriptor; spike rate not age-informative; used as confound regressor | Spikes, rhythmic periodicity |
| `pathological_flagging` | 60 s | ∼18 | Binary flags for specialist patterns used as state covariates | Triphasic, GRDA, polyspikes |
| `sigma_band` | 30 s | 117 | Established brain-age marker in sleep EEG; coma spindles reflect thalamocortical circuit integrity | — |
| `wavelet` | 60 s | ∼4 | Hemispheric coherence and wavelet entropy | — |
| **Tier-2 primary families (selected)** | | ∼1,128 | See Table A.7 for full module totals | |

### Feature Count Breakdown: Extractor Modules

The 3,423 post-deduplication features (Profile B/D) originate from eight extractor modules. Table A.7 lists the pre-deduplication design-specification counts from the February 2026 pipeline specification; 497 features were removed in the deduplication pass that identified exact-duplicate column names arising where monopolar and bipolar computations produced identically named outputs.

**Table A.7:** Feature count by extractor module. Pre-deduplication counts are from the February 2026 pipeline specification (v3.0). A deduplication pass removed 497 features with duplicate column names arising from overlapping monopolar and bipolar computations, giving a post-deduplication total of 3,423 for Profile B/D. Profile A/C/E (19 monopolar channels only) extracts 1,578 features; the 1,845-feature increment in Profile B/D arises from computing selected extractors across the ten bipolar channel pairs listed in Section A.1.3.

| Module | Primary outputs | Pre-dedup count |
|---|---|---|
| `core` | Spectral, aperiodic (specparam), temporal, Hjorth, nonlinear | 1,460 |
| `wavelet` | DWT energies, coefficients, entropy (Daubechies-4) | 639 |
| `burst` | BSR, IBI, burst amplitude/duration, HFO rate | 380 |
| `riemannian` | Covariance geometry, log-Euclidean distances (6 bands) | 1,158 |
| `aeeg` | aEEG margins, bandwidth, mean, Hellström-Westas grade | 121 |
| `connectivity` | PLI, wPLI, ImCoh, coherence (interhemispheric & fronto-posterior) | 75 |
| `additional` | BIS approximation, beta ratio, synch_fast_slow | 59 |
| `pathological` | Spike rate, periodicity, sharp transients, ACNS flags | 28 |
| **Pre-dedup total (Profile B/D)** | | **3,920** |
| *Deduplication removed* | *Duplicate names from monopolar/bipolar overlap* | *−497* |
| **Post-dedup total (Profile B/D)** | | **3,423** |
| **Profile A/C/E (monopolar only)** | | **1,578** |
| **Bipolar increment (10 pairs, Profile B/D)** | | **1,845** |

### A.4.1 Spectral and Aperiodic Traits

Spectral features are computed per 60-second window per channel using Welch's method (Hann window, 50% overlap), with a frequency resolution of 0.25 Hz at 256 Hz sampling rate. After patient-level aggregation (Section A.5), the spectral family contributes approximately 300 dimensions to the final feature matrix.

**Table A.8:** Spectral and aperiodic features extracted per channel per 60-second window. Aperiodic features require `specparam` ≥v2.0; failed fits (NMAR) are retained as a feature value in their own right.

| Feature | Band / range | Unit | Rationale |
|---|---|---|---|
| *Absolute band power (Welch PSD)* | | | |
| `delta_power` | 0.5–4 Hz | $\mu V^2$ | Primary marker of cortical slowing; elevated in coma and acute brain injury |
| `theta_power` | 4–8 Hz | $\mu V^2$ | Drowsiness, encephalopathy, metabolic disturbance |
| `alpha_power` | 8–13 Hz | $\mu V^2$ | Resting wakefulness; decreases with age (≈1 Hz/decade) and sedation |
| `beta_power` | 13–30 Hz | $\mu V^2$ | Alertness; elevated by benzodiazepines |
| `gamma_power` | 30–50 Hz | $\mu V^2$ | Cognitive processing; frequently contaminated by EMG in ICU |
| `sigma_power` | 11–16 Hz | $\mu V^2$ | Coma-spindle activity; marker of intact thalamocortical circuitry |
| *Relative band power (fraction of 0.5–45 Hz total; dimensionless)* | | | |
| `*_power_rel` | (same bands) | — | Amplitude-normalised version of each absolute power; more robust to inter-patient amplitude variation |
| *Clinical encephalopathy ratios* | | | |
| `delta_alpha_ratio` | $\delta/\alpha$ | — | Encephalopathy marker; elevated ratio indicates progressive cortical slowing |

*Continued from previous page*

| Feature | Band / range | Unit | Rationale |
|---|---|---|---|
| `theta_alpha_ratio` | $\theta/\alpha$ | — | Complementary encephalopathy marker |
| *Spectral shape features* | | | |
| `peak_alpha_freq` | 8–13 Hz | Hz | Individual alpha frequency; known to decrease ≈1 Hz per decade after age 20 |
| `sef95` | 0.5–45 Hz | Hz | Spectral edge frequency: frequency below which 95% of spectral power lies |
| `spectral_entropy` | Welch PSD | bits | Shannon entropy of normalised PSD; low = peaked spectrum, high = broadband |
| `spectral_centroid` | 0.5–45 Hz | Hz | Power-weighted mean frequency; decreases with cortical slowing |
| *Aperiodic / 1/f features (specparam; NMAR when fitting fails)* | | | |
| `specparam_exponent` | 1–45 Hz fit | — | $1/f$ slope; *decreases* (flattens) with healthy ageing; artificially elevated (steepened) by sedation and severe cortical suppression; reflects excitation/inhibition balance |
| `specparam_offset` | 1–45 Hz fit | $\log(\mu V^2/Hz)$ | Broadband power level (y-intercept of the aperiodic fit) |
| `specparam_r_squared` | — | — | Goodness of fit; values $< 0.8$ retained as NMAR injury marker |
| `*_power_corrected` | (all five bands) | $\mu V^2$ | Band power after subtracting the aperiodic background; isolates true oscillatory activity |
| *Amplitude-integrated EEG (aEEG; bandpass 2–15 Hz, per clinical standard)* | | | |
| `aeeg_upper` | P95 envelope | $\mu V$ | Burst amplitude estimate |
| `aeeg_lower` | P5 envelope | $\mu V$ | Suppression depth estimate |

*Continued from previous page*

| Feature | Band / range | Unit | Rationale |
|---|---|---|---|
| `aeeg_bandwidth` | upper − lower | $\mu$V | Continuity index of cortical activity |
| `aeeg_mean` | — | $\mu$V | Overall signal energy in aEEG band |
| *BIS approximation features* | | | |
| `bis_beta_ratio` | $\log_{10}(P_{30\text{–}47\,\mathrm{Hz}}/P_{11\text{–}20\,\mathrm{Hz}})$ | — | BIS correlate; higher values correspond to lighter sedation |
| `synch_fast_slow` | envelope correlation | — | Fast–slow band coupling; anaesthesia-depth correlate |
| `bis_approximation` | empirical 0–100 scale | — | BIS surrogate; not FDA-validated but provides a single-dimensional consciousness estimate |

## A.4.2 Temporal and Complexity Traits

Time-domain, Hjorth, burst-suppression, and wavelet features are extracted per 60-second window. The burst-suppression features use a minimum burst/suppression duration of 0.5 s (ACNS criteria) and an RMS amplitude threshold of 10 $\mu$V to distinguish cortical bursts from suppressed background.

**Table A.9:** Temporal, complexity, burst-suppression, sigma-band, and wavelet features extracted per channel per 60-second window. Three complexity features (sample entropy, DFA exponent, Higuchi FD) were re-enabled with signal-length caps; approximate entropy remains disabled.

| Feature | Description | Unit | Rationale |
|---|---|---|---|
| *Time-domain amplitude statistics* | | | |
| `mean_amplitude` | Mean $\lvert x(t)\rvert$ | $\mu$V | Overall signal strength |
| `std_amplitude` | Standard deviation | $\mu$V | Signal variability |
| `rms_amplitude` | Root mean square | $\mu$V | Signal energy |
| `peak_to_peak` | Max − Min | $\mu$V | Dynamic range |
| `kurtosis` | 4th standardised moment | — | Peakedness; elevated in burst-suppression |

*Continued from previous page*

| Feature | Description | Unit | Rationale |
|---|---|---|---|
| `skewness` | 3rd standardised moment | — | Amplitude asymmetry |
| `zero_crossing_rate` | Crossings per second | Hz | Crude dominant frequency estimate |
| `line_length` | $\sum \lvert\Delta x\rvert \,/\, T$ | $\mu$V/s | Signal complexity; sensitive to high-frequency content |
| *Hjorth parameters (time-domain complexity)* | | | |
| `hjorth_activity` | $\mathrm{var}(x)$ | $\mu\mathrm{V}^2$ | Signal power in time domain |
| `hjorth_mobility` | $\sqrt{\mathrm{var}(\dot{x})/\mathrm{var}(x)}$ | Hz | Proportional to dominant frequency; equivalent to $2\pi f$ for pure sinusoid |
| `hjorth_complexity` | $\mathrm{mobility}(\ddot{x})$ / $\mathrm{mobility}(\dot{x})$ | — | Deviation from sinusoid; $= 1.0$ for pure sine; elevated by multi-frequency content |
| *Burst-suppression quantification (global, across all channels)* | | | |
| `burst_suppression_ratio` | Suppressed fraction of window | 0–1 | Primary coma severity and prognosis marker |
| `inter_burst_interval` | Mean IBI | s | Regularity of the burst rhythm |
| `burst_rate` | Bursts per minute | $\mathrm{min}^{-1}$ | Rate of cortical excitability recovery |
| `burst_amplitude_mean` | Mean burst RMS amplitude | $\mu$V | Burst signal strength |
| `burst_duration_mean` | Mean burst duration | s | Burst morphology |
| `perm_entropy_burst` | Permutation entropy of burst segments | bits | Brain-age marker within bursts (cf. Kratzer et al. 2020) |
| `intra_burst_delta` | $\delta$-band power within bursts | $\mu\mathrm{V}^2$ | Spectral content of cortical bursts |
| *Sigma-band / coma-spindle features (11–16 Hz)* | | | |
| `sigma_burst_rate` | Spindle-like events per minute | $\mathrm{min}^{-1}$ | Thalamocortical integrity; associated with favourable neurological outcome |

*Continued from previous page*

| Feature | Description | Unit | Rationale |
|---|---|---|---|
| `sigma_burst_count` | Raw event count per window | — | Spindle burden |
| `sigma_burst_duration_mean` | Mean event duration | s | Spindle morphology |
| `sigma_burst_amplitude_mean` | Mean envelope peak amplitude | $\mu$V | Spindle strength |
| *Wavelet features (Daubechies-4 DWT, per clinical band)* | | | |
| `wavelet_energy_*` | Per-band DWT energy (5 bands) | J | Time-frequency localised energy; complements Welch PSD for non-stationary signals |
| `wavelet_entropy` | Shannon entropy of DWT coefficients | bits | Time-frequency spectral complexity |
| `wavelet_energy_rel_*` | Relative band energies (5 bands) | — | Amplitude-normalised wavelet distribution |
| `wavelet_coeff_*` | Mean, std, skewness, kurtosis per band | — | DWT coefficient distribution shape |
| *Re-enabled complexity features (computational mitigations applied; see Section A.4.4)* | | | |
| `sample_entropy` | SampEn, $m$=2, $r$=0.2$\sigma$ | — | Re-enabled: $O(N^2)$ mitigated; signal capped at first 10 s (2 560 samples at 256 Hz); `antropy` backend |
| `dfa_exponent` | Detrended Fluctuation Analysis | — | Re-enabled: signal capped at first 30 s; `antropy` backend |
| `higuchi_fd` | Higuchi fractal dimension | — | Re-enabled: signal capped at first 30 s; `neurokit2` backend |
| *Disabled features (no mitigation; see Section A.4.4)* | | | |
| `approx_entropy` | ApEn, $m$=2 | — | Disabled: similar $O(N^2)$ cost to SampEn; no mitigation applied |

### A.4.3 Pathological and ACNS-Aligned Features

The pipeline includes numerical proxies for a subset of the clinical EEG patterns defined by the American Clinical Neurophysiology Society (ACNS) 2021 Critical Care EEG Terminology [89]. These features are extracted by dedicated modules (`pathological_features.py`, `pathological_flagging.py`) and carry the `patho_*` prefix in the feature matrix.

**Scope and limitations.** Detection relies on amplitude and slope thresholds applied to preprocessed single-channel windows rather than gold-standard expert visual review. Patterns that require multi-channel spatial context for reliable identification — including triphasic waves, generalised periodic discharges at low amplitude, and polyspike-wave complexes — are excluded from automated extraction (see Section A.4.4). The features below are best interpreted as quantitative state descriptors that *correlate with* the corresponding ACNS categories; they do not constitute a clinical diagnosis.

**Table A.10:** ACNS-aligned pathological features extracted per channel per 60-second window. Detection uses amplitude/slope thresholds; results are not equivalent to expert visual review. ACNS category refers to the nearest category in the 2021 Critical Care EEG Terminology [89].

| Feature | Description | Unit | ACNS category / rationale |
|---|---|---|---|
| *Spike activity* | | | |
| `spike_rate_per_min` | Detected spikes per minute | $\text{min}^{-1}$ | Epileptiform discharges / spikes; elevated in post-anoxic status epilepticus |
| `spike_amplitude_mean/std` | Mean and SD of spike amplitudes | $\mu$V | Spike morphology |
| `spike_duration_mean/std` | Mean and SD of spike widths | ms | Polarity and morphology indicator |
| `interspike_interval_mean/std` | Mean and SD of ISI | s | Discharge regularity |
| `spike_regularity_index` | $1/\text{CV}_{\text{ISI}}$ | — | High value indicates rhythmic (periodic) discharges |
| *Periodic discharge pattern* | | | |
| `periodicity_index` | Spectral periodicity score | — | Periodic discharges (GPD, LPD, BIPD analogue) |

*Continued from previous page*

| Feature | Description | Unit | ACNS category / rationale |
|---|---|---|---|
| `dominant_period_sec` | Period of dominant discharge | s | Discharge frequency proxy |
| `periodic_amplitude_cv` | CV of periodic amplitude amplitude | — | Morphological stability of discharges |
| *Sharp transients* | | | |
| `sharp_transient_rate` | Sharp events per minute | $\text{min}^{-1}$ | Epileptiform sharp waves / interictal discharges |
| `max_slope_uv_per_ms` | Peak leading-edge slope | $\mu$V/ms | Transient sharpness indicator |
| `slope_asymmetry` | Rise/fall slope ratio | — | Classic sharp-wave asymmetry criterion |
| *High-frequency oscillations (HFO)* | | | |
| `ripple_rate_per_min` | Events in 80–250 Hz band | $\text{min}^{-1}$ | Ripple-band HFOs; prognostic in hypoxic-ischaemic encephalopathy |
| `fast_ripple_rate_per_min` | Events in 250–500 Hz band | $\text{min}^{-1}$ | Fast ripples; indicative of severe injury |
| *Pathological state flags (`pathological_flagging` module; aggregated as fraction of windows positive)* | | | |
| `seizure_detected` | Binary seizure flag | 0/1 | Seizures |
| `burst_suppression_present` | Binary burst-suppression flag | 0/1 | Suppression patterns |
| `sigma_spindle_flag` | Binary coma-spindle flag | 0/1 | Sigma-band spindle-like event |
| `ecg_contamination_flag` | Binary ECG-artefact flag | 0/1 | Residual cardiac contamination |

### A.4.4 Feasibility Constraints and Heuristic Limits

Three categories of constraint limit what can feasibly and reliably be extracted from this cohort.

**Computational feasibility.** Three complexity metrics that were initially prohibitive under full 60-second window computation — sample entropy ($O(N^2)$; profiled at 1318 s per patient), DFA exponent (≈90 s per patient), and Higuchi fractal dimension (≈10 s per patient) — have been re-enabled in the production pipeline with signal-length caps. Sample entropy is capped at the first 10 s of each window (2 560 samples at 256 Hz), reducing the per-patient cost below 1 s via the `antropy` C-backed implementation; the 10 s subsample is sufficient for reliable SampEn estimation ($N \geq 1000$, $m$=2). DFA exponent and Higuchi fractal dimension are both capped at the first 30 s per window using the `antropy` and `neurokit2` backends respectively. Approximate entropy (`approx_entropy`) remains disabled: it carries the same $O(N^2)$ profile as SampEn without a well-characterised minimum-$N$ convergence criterion that would justify a fixed cap.

**Polyspike and triphasic wave detection.** No validated open-source library existed (as of 2026) for reliable automatic detection of polyspikes or triphasic waves in comatose EEG. Libraries validated on healthy or lightly sedated EEG — including YASA, MNE, and mne-hfo — fail to generalise to the globally synchronised, low-amplitude burst-suppression morphology typical of post-cardiac-arrest coma. The pipeline therefore restricts pathological feature extraction to: (a) burst-suppression quantification (BSR, IBI, burst amplitude and duration; validated for coma EEG and listed in Table A.9), (b) spike rate and periodicity metrics derived from amplitude and slope thresholds, and (c) sigma-band spindle-like event detection for coma-spindle identification.

**Connectivity features: selective computation.** Full pairwise inter-channel coherence and phase-locking value (PLV) across all $C(19, 2) = 171$ electrode pairs would add approximately 855 coherence and 1026 PLV features. Given the already high-dimensional feature space and HPC runtime budget, connectivity computation is restricted to inter-hemispheric coherence (8 pairs across 5 frequency bands), fronto-posterior coherence (4 pairs across 5 bands), and global mean and standard deviation per band. Full pairwise connectivity matrices are deferred to future work with dedicated dimensionality reduction.

## A.5 Feature-Adaptive Hierarchical Aggregation

The reduction of the temporal window sequence to a stable patient-level biometric is governed by a hierarchical protocol.

**Table A.11:** Feature-adaptive aggregation classes. Each base feature is assigned to one class based on its empirical distribution (Table A.11).

| Class | Label | Statistics | Mult. |
|---|---|---|---|
| 266-A | Continuous | Median, IQR, skewness, 5th/95th percentile | $\times 5$ |
| 267-B | Bimodal | Median, IQR, skewness, bimodal split ratio | $\times 4$ |
| 268-C | Binary | Proportion, count | $\times 2$ |
| 269-D | Count | Mean, max | $\times 2$ |

## A.6 Feature Domain Clusters

The 3,423 base features are grouped into 16 functional domains. Table A.12 details the domain composition, aggregation class assignment, median intra-class correlation, and the proportion of features classified as bimodal or temporally trending. Figure A.6 shows the quality metric distributions across the main domain clusters.

**Table A.12:** Feature domain summary ($N = 3{,}423$ base features). Aggregation classes: A Continuous, B Bimodal, C Binary, D Count. Median ICC computed from EDA-010-B ($N = 366$ patients). Burst-Suppression includes channel-wise and global descriptors (BSR, IBI, burst morphology, suppression depth), not only binary flags. % Bimodal and % Trending are proportions of features in each domain flagged by the distributional classifier.

| Domain | *N* feat. | *N* prefix | Agg class | Med. ICC | % Bimodal | % Trending |
|---|---|---|---|---|---|---|
| Band Power | 720 | 144 | B | 0.30 | 72.9 | 80.8 |
| Burst-Suppression | 646 | 34 | B | 0.30 | 79.7 | 24.8 |
| Aperiodic | 495 | 99 | A | 0.60 | 0.0 | 95.8 |
| Other | 265 | 13 | A | 0.50 | 25.7 | 82.6 |
| aEEG | 229 | 13 | A | 0.70 | 3.1 | 99.6 |
| Wavelet | 185 | 185 | A | 0.00 | 0.0 | 0.0 |
| Spectral Shape | 180 | 36 | B | 0.50 | 56.7 | 98.3 |
| Spectral Ratio | 172 | 28 | B | 0.10 | 54.1 | 100.0 |
| Segment Counts | 166 | 22 | D | 0.40 | 47.6 | 31.3 |
| Signal Stats | 135 | 27 | A | 0.50 | 15.6 | 94.1 |
| Complexity | 90 | 18 | A | 0.60 | 1.1 | 100.0 |
| Pattern Flags | 49 | 49 | C | 0.30 | 83.7 | 12.2 |
| BIS | 39 | 3 | A | 0.50 | 0.0 | 100.0 |
| Connectivity | 39 | 3 | A | 0.50 | 5.1 | 15.4 |
| Regional Avg. | 8 | 8 | A | 0.60 | 0.0 | 100.0 |
| ACNS | 5 | 5 | B | 0.50 | 80.0 | 80.0 |
| **Total** | **3,423** | **687** | | | | |

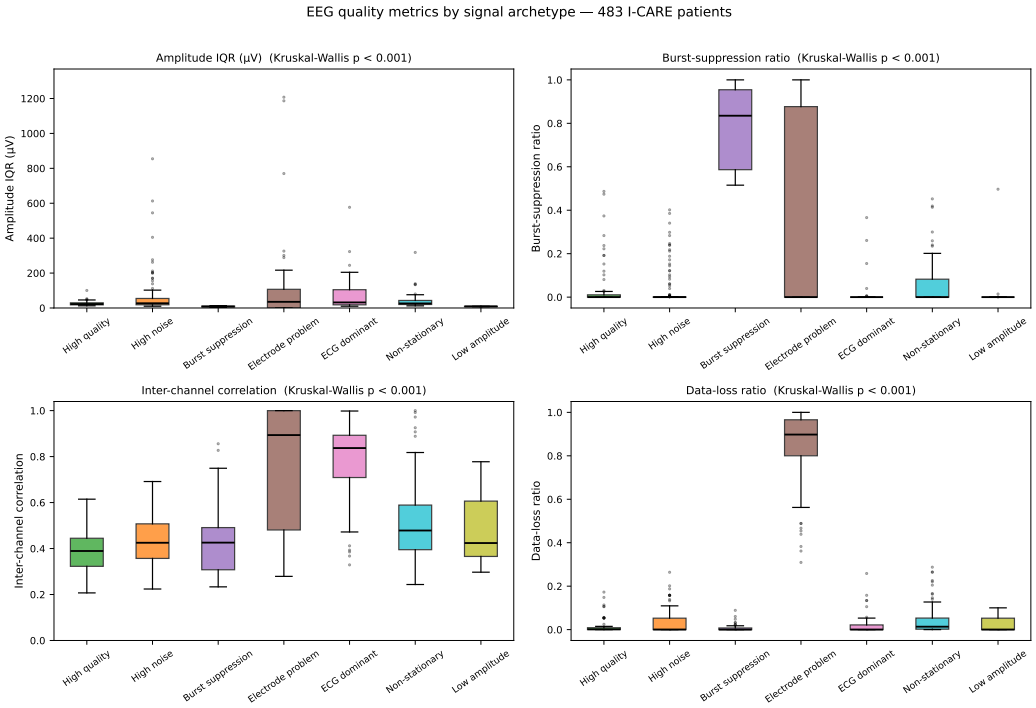



**Figure A.6:** Distribution of quality metrics across patient clusters. Figure A.6 shows the separation of technical noise from physiological variance.

## A.7 Feature Distributions and EDA

This section collects the complete suite of exploratory data analysis (EDA) figures generated during the thesis project. These materials substantiate the design decisions described in Sections 4.4, 4.4.1 and 4.4.2 and cover four complementary perspectives: raw feature visualisations on real ICU EEG, within- and between-patient variability, temporal feature drift, feature–outcome and feature–demographic correlations, and feature domain summary dashboards.

The majority of analyses (EDA-010-B) were conducted on the Phase 1 development split using Profile B preprocessing (spectral cleaning with ECG-ICA, $n = 366$ patients). Profile B was chosen because it represents the most complete preprocessing chain available locally; full-corpus results for all profiles ($n = 604$ patients) reside on the GWDG/Kisski cluster (`/mnt/vast-kisski/.../07-data/03-feature-extraction-preprocessed/`). Profile C produced no local feature extractions (feature extraction for Profile C was run exclusively on the HPC cluster; see Section A.7.6), so figures that span multiple profiles show Profile C without feature data. The primary scripts are `eda010_server_variants.py`, `generate_aggregation_figures.py`,

and `fig_features_real_data.py` in `04-code/03-eda/` and `04-code/05-visualization/`. Complementary interactive exploration was conducted in `phase1_eda_window_features3.ipynb` (`04-code/00-notebooks/`) and `interactive_metadata_analysis.ipynb` (`04-code/08-artifacts/`); the latter provides a real-time mediation explorer with bootstrap confidence intervals for clinical metadata pathways. Pipeline-level quality figures (signal-and-mask viewer, PSD comparison, mask timeline) are documented separately in Section C.4.3.

### A.7.1 Feature Visualisations on Real ICU EEG

Figure A.7 illustrates the four families of handcrafted features extracted from patient 0284 (I-CARE training cohort), processed under Profile E (full preprocessing pipeline: bandpass, re-referencing to average, ECG-ICA, quality masking). It covers *all 60-second windows across all recording segments* for this patient concatenated in temporal order; it is not restricted to the first hour of recording (which coincides with TTM administration and maximum sedation and is therefore the most confounded period). Patient 0284 was chosen as a representative I-CARE training case with a full-length recording ($\approx$32 hours) and good feature completeness. This visualisation supports RO-1 and RO-2: it confirms that the handcrafted feature extractor produces systematic, interpretable variation in spectral power, aEEG, and temporal complexity across the full ICU monitoring window — a prerequisite for brain-age regression and outcome prediction.

### A.7.2 Intra- and Inter-Patient Feature Variability

Within-patient consistency was quantified by ICC(3,1) across 60-second windows (experiment EDA-010-B, Profile B, Phase 1 development split, $n = 366$ patients). Figure A.8 shows the ICC categorisation broken down by feature domain ($N = 3{,}423$ features): 60.3% fall in the poor range (ICC $< 0.4$), 38.4% moderate ($0.4$–$0.75$), and only 0.8% ($n = 27$) achieve good ICC ($> 0.75$). These proportions are prerequisite evidence for all outcome-prediction and brain-age research objectives (RO-1–RO-3): only features with adequate within-patient stability can produce meaningful patient-level summaries. They directly motivated the trajectory-aware aggregation strategy (Section A.5). Figure A.9 extends this with outcome-stratified IQR distributions (EDA-010-B, Profile B), revealing which features show greater spread within the poor-outcome class — an important complement to point-estimate discriminability for RO-1.

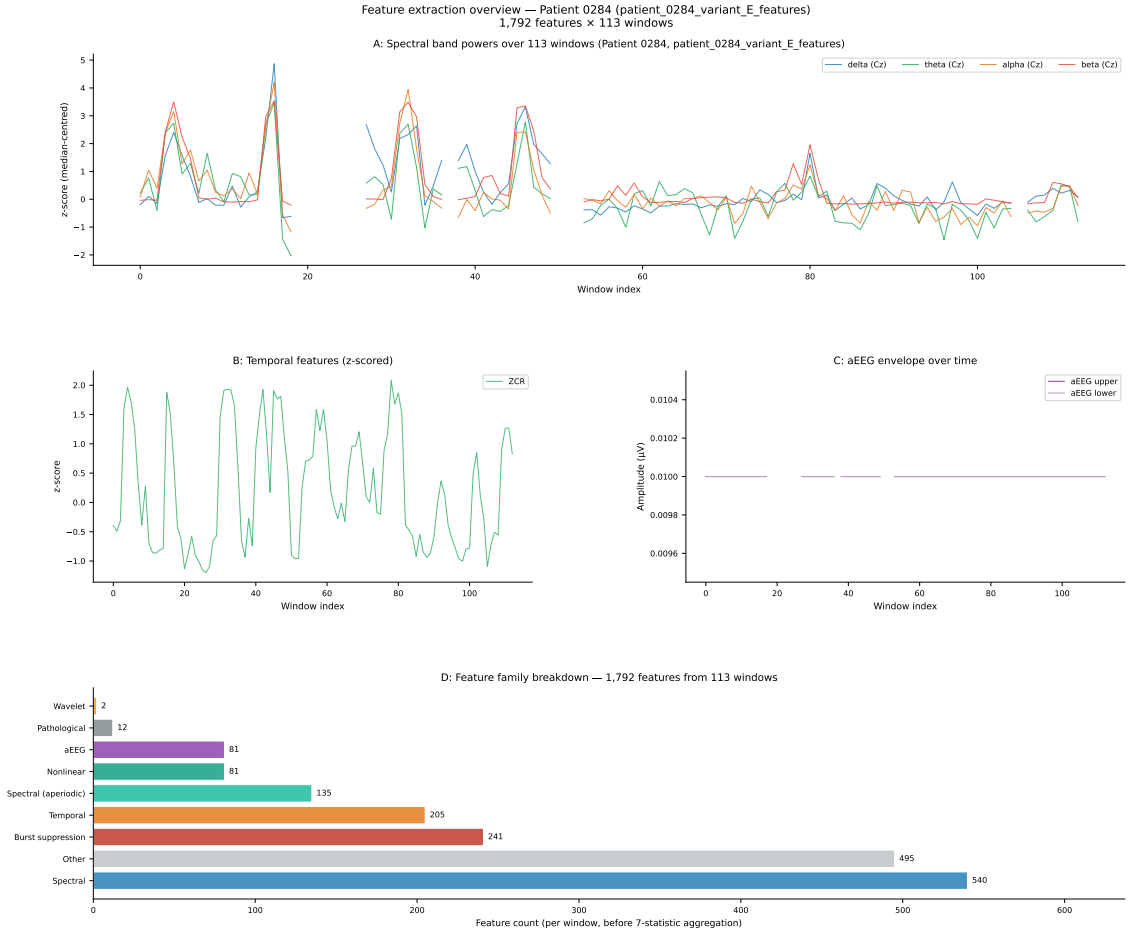



**Figure A.7:** Four feature families extracted from patient 0284 (I-CARE training cohort), Profile E preprocessing. All 60-second windows are shown in temporal order across all recording segments (not restricted to the first hour). *Top:* Spectral band-power time series (delta, theta, alpha, beta; channel Cz). *Second:* aEEG amplitude envelope. *Third:* Temporal complexity features (Hjorth mobility, complexity, line-length, zero-crossing rate). *Bottom:* Count of successfully extracted feature families per window. Systematic temporal variation is visible in all panels, confirming that the feature extractor produces clinically interpretable outputs (RO-1, RO-2).

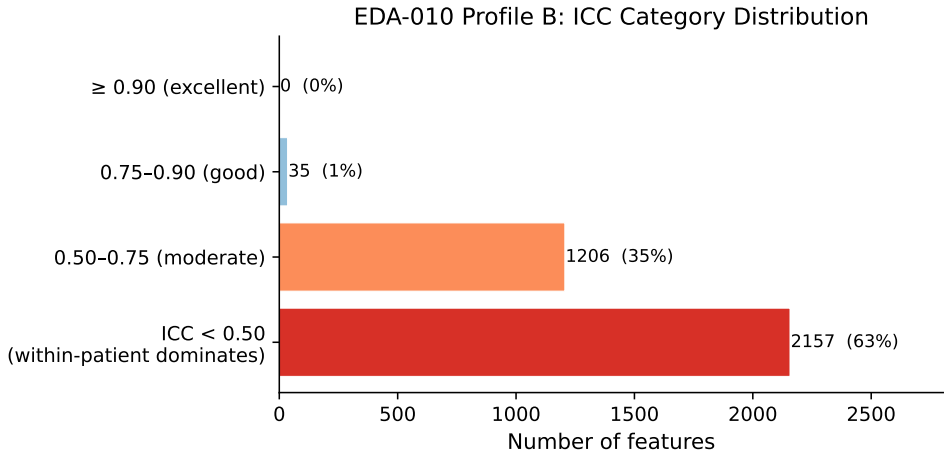



**Figure A.8:** ICC categorisation (poor / moderate / good) by feature domain. **Experiment:** EDA-010-B. **Data:** Profile B preprocessing (spectral cleaning + ECG-ICA), Phase 1 development split, $N = 3{,}423$ features, $n = 366$ patients. The dominance of the poor category (60.3%) confirms that point-estimate aggregation is insufficient and that trajectory statistics (slope, IQR, percentiles) are required to capture meaningful between-patient differences. These proportions are prerequisite for RO-1–RO-3.

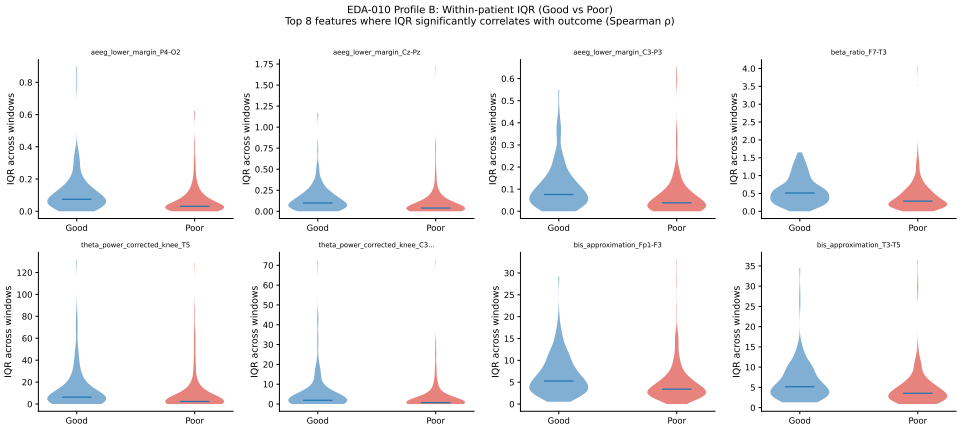



**Figure A.9:** Interquartile range (IQR) of patient-level feature distributions stratified by neurological outcome (Good: CPC 1–2; Poor: CPC 3–5). **Experiment:** EDA-010-B. **Data:** Profile B, Phase 1 development split, $n = 366$ patients. Features with larger IQR spread in the poor-outcome class represent heterogeneous clinical trajectories and motivate variance-based aggregation statistics. Supports RO-1 (neurological outcome prediction).

### A.7.3 Temporal Feature Drift — Extended Analysis

Figure 4.9 in Section 4.4.1 shows aggregate temporal drift across all features. Figures A.10 and A.11 extend this with outcome-stratified views (experiment EDA-010-B, Profile B, Phase 1 development split, $n = 366$ patients). Both figures cover *all recording segments* spanning the first 36 hours of post-arrest monitoring; they are not restricted to any single recording segment. Figure A.10 shows individual feature trajectories coloured by CPC outcome, revealing that divergence between good and poor outcome groups begins within the first 12–24 hours after arrest. Figure A.11 shows group-level temporal evolution with bootstrap confidence bands, making the TTM-confounded suppression phase in the first 24 hours visible for the poor-outcome group (Section 2.1). These findings motivated the 24-hour holdout and the trajectory-feature aggregation design, and provide empirical support for RO-1.

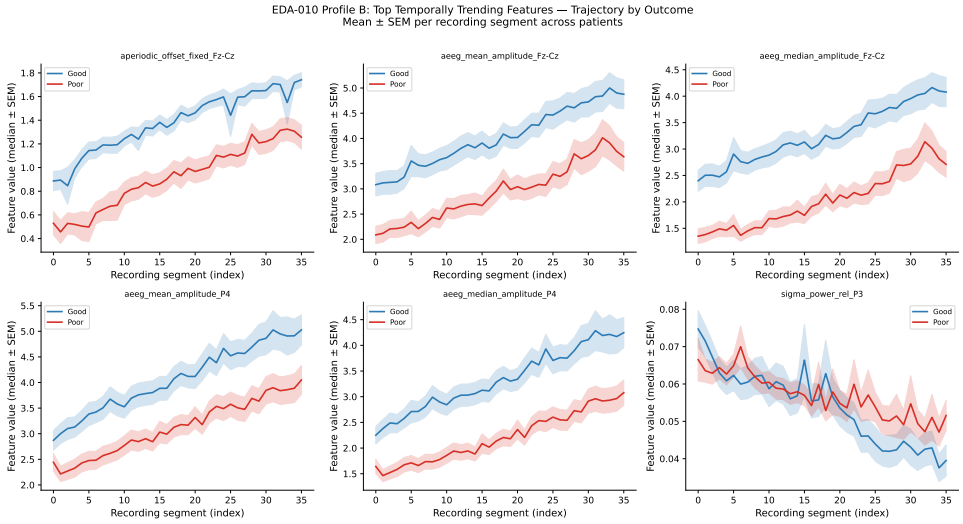



**Figure A.10:** Feature trajectories across all recording segments spanning the first 36 hours of post-arrest monitoring, stratified by neurological outcome (Good: CPC 1–2; Poor: CPC 3–5). **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. Early divergence — often within the first 12–24 hours — is visible in spectral and complexity features. The initial suppression phase in the poor-outcome group reflects TTM and sedation confounders. Supports RO-1.

### A.7.4 Feature–Outcome and Feature–Demographic Correlations

These analyses do not appear in the main thesis chapters but provide the statistical foundation for feature selection and confounder identification. All EDA-010-B figures use Profile B preprocessing, Phase 1 development split ($n = 366$ patients). The

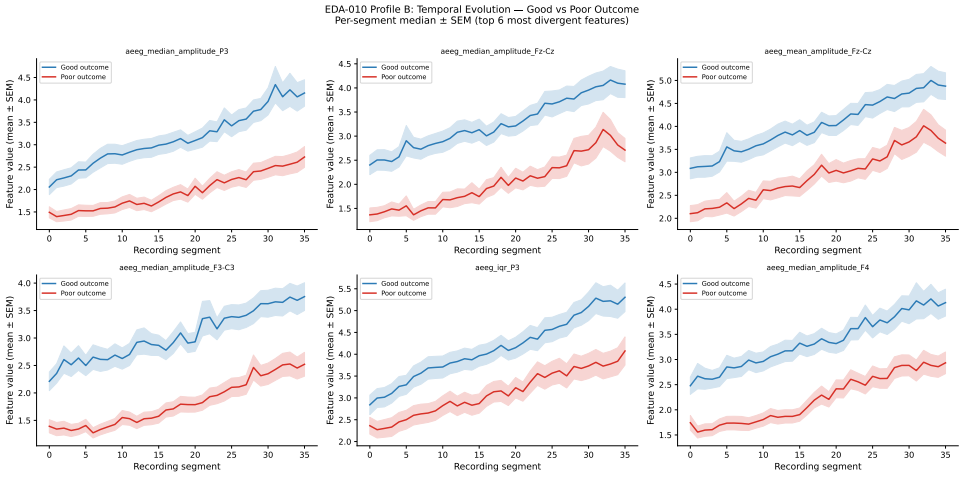



**Figure A.11:** Group-level temporal evolution of key neurophysiological features by outcome, with bootstrap confidence bands. **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. Covers all recording segments across the 36-hour monitoring window (not a single segment). The suppression-phase plateau in the poor-outcome group during the first 24 hours reflects TTM-induced metabolic downshift (Section 4.4.1). Motivates the TTM-aware feature design and supports RO-1.

age–sex correlation figure uses the Phase 1 training split ($n = 313$ patients, patient-level median aggregation over all recording segments, Profile A baseline preprocessing; `phase1_preprocessed_window_features_PHASE1TRAIN.parquet`).

Figure A.12 shows Spearman $\rho$ between each feature and binary neurological outcome (RO-1). Figure A.13 shows the KS-test statistic for Good versus Poor CPC separation; 168 of the 3,423 Profile B features achieved better discriminability with IQR aggregation than with median aggregation, motivating the multi-statistic scheme. Figure A.14 shows the top features by Spearman $\rho$ with age and sex (RO-1). Figure A.15 shows a full feature–demographic heatmap underpinning the confounder-stratified design.

### A.7.5 Feature Domain Dashboard

Figures A.16 to A.18 summarise the feature domain structure and the bimodality-driven aggregation design (experiment EDA-010-B, Profile B, Phase 1 development split, $n = 366$ patients). Figure A.16 shows the proportion of bimodal features by domain (IQR-split criterion: $> 50\%$ of patients); 1,458 features met this criterion, driving the decision to retain IQR as a trajectory summary. Here, "Burst-Suppression" denotes the engineered descriptor family (e.g., BSR, burst rate, IBI statistics), not a direct binary bedside diagnosis label. Figure A.17 shows the demographic correlation heatmap on aggregated patient-level

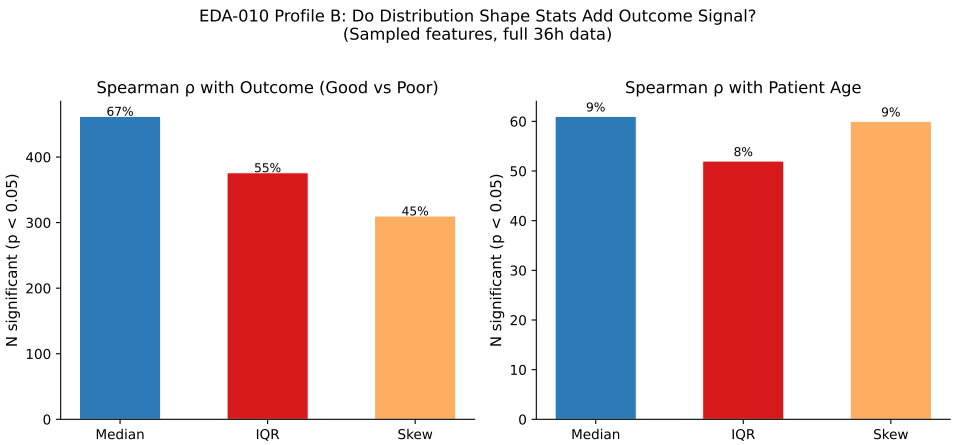



**Figure A.12:** Spearman $\rho$ between each feature and binary neurological outcome (Good: CPC 1–2 vs. Poor: CPC 3–5), grouped by feature domain. **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. The strongest associations are concentrated in spectral power and burst-suppression descriptors. Directly supports RO-1 (neurological outcome prediction) and the Phase 1 feature selection described in Section 4.5.

features, confirming that the aggregation step does not inflate confounding associations. Figure A.18 provides the full domain summary dashboard with feature class distribution, bimodality rates, and ICC summaries per domain, validating the taxonomy in Section A.6. These figures collectively support all ROs dependent on patient-level feature quality (RO-1–RO-3).

### A.7.6 Segment Duration and Valid-Window Analysis (EDA-011)

EDA-011 characterised the recording duration and valid-window structure across Profiles A, B, C, and E on the full cohort ($n = 604$ patients). The median total recording duration is 32.3 hours, reflecting the 36-hour I-CARE monitoring window. Recording duration and quality data are available for all profiles.

Figure A.19 shows total recording duration by profile; the median is 32.3 hours (Profiles A and E), reflecting the 36-hour I-CARE monitoring window. The companion EDA-011 quality-yield analysis (reported in text) showed a median valid-window ratio of 0.986 for Profiles A and E, with 76 patients below 50% quality retention. Together, these results validated the 60-second non-overlapping window scheme and guided the minimum-segment threshold in feature extraction, supporting the preprocessing design for all ROs.

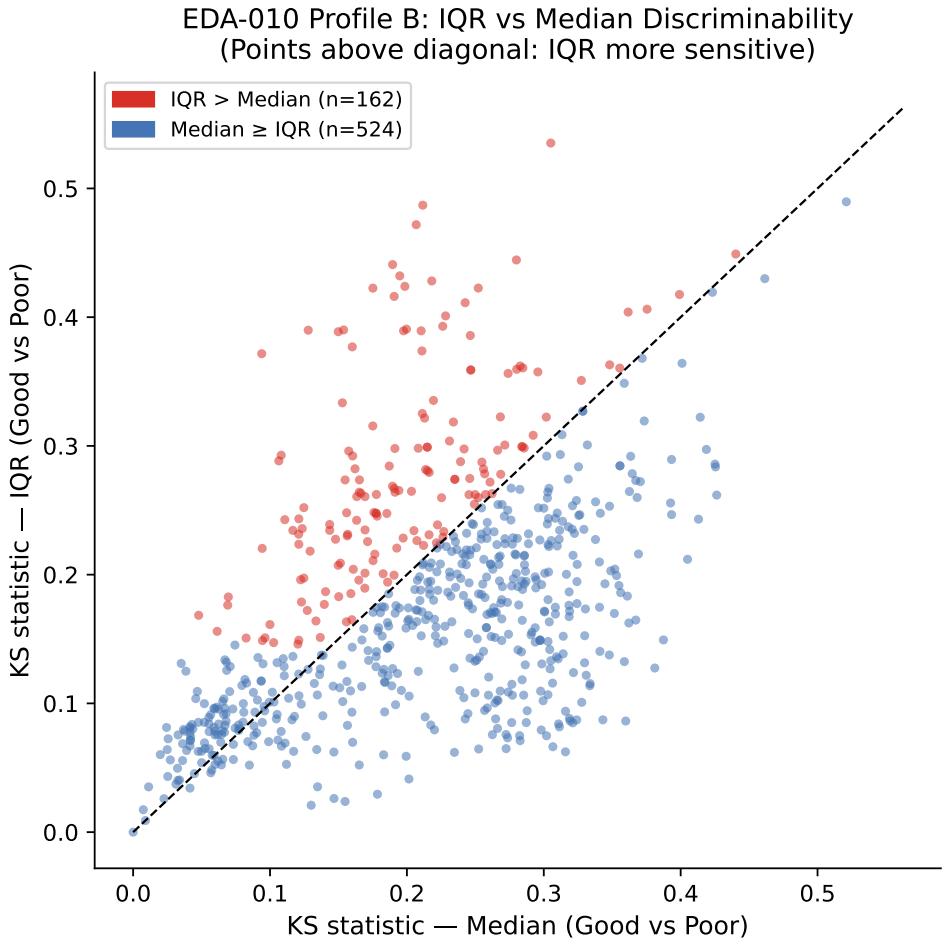



**Figure A.13:** Kolmogorov–Smirnov (KS) test statistic for each feature comparing Good CPC (1–2) and Poor CPC (3–5) patient distributions. **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. KS $> 0.3$ indicates moderate distributional separation. Of the 3,423 features, 168 achieved better discriminability with IQR aggregation than median aggregation, motivating the multi-statistic scheme. This does not imply near-linear class separation at the patient level; it indicates that spread-sensitive summaries capture additional heterogeneity that point estimates miss. Supports RO-1.

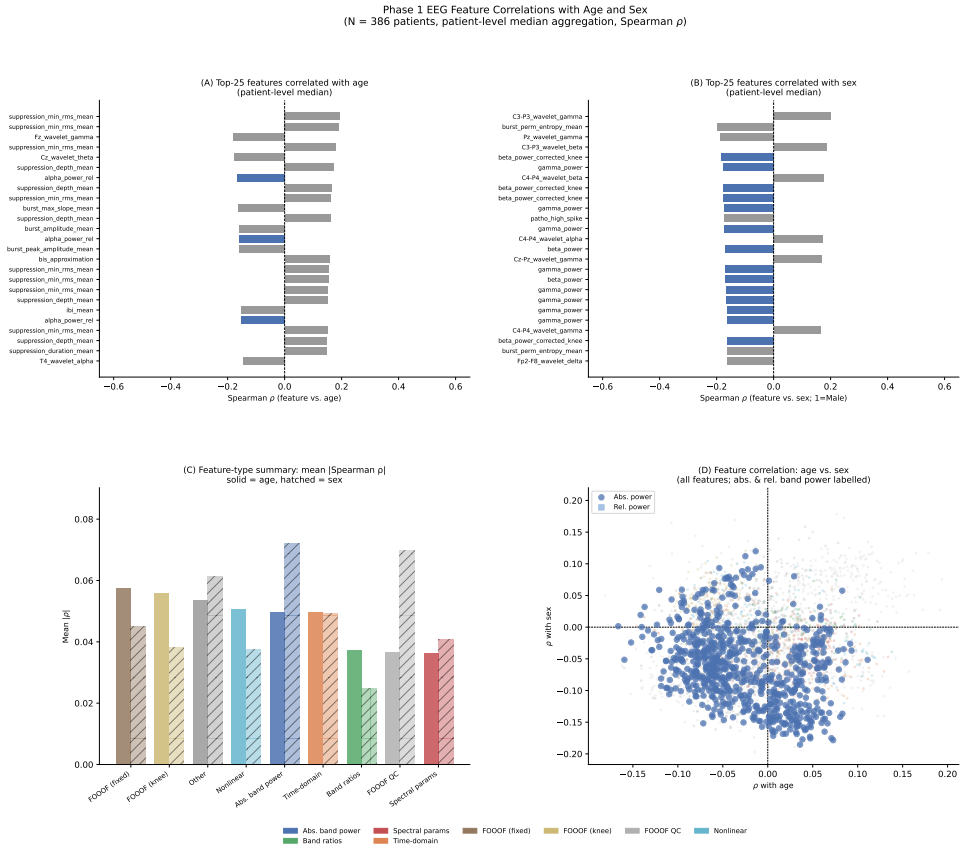



**Figure A.14:** Feature correlations with chronological age and biological sex. Panels A and B rank the strongest individual features by Spearman $\rho$ for age and sex respectively; Panel C summarises the mean absolute correlation by feature family; and Panel D maps the joint age-versus-sex correlation structure across the full feature set. **Data:** Phase 1 training split, patient-level median aggregation under Profile A baseline preprocessing. The figure is used here as an EDA reference for demographic confounding: age-related effects are concentrated in a limited subset of domains, while sex-related effects are weaker and only partially overlap with the age-associated feature families.

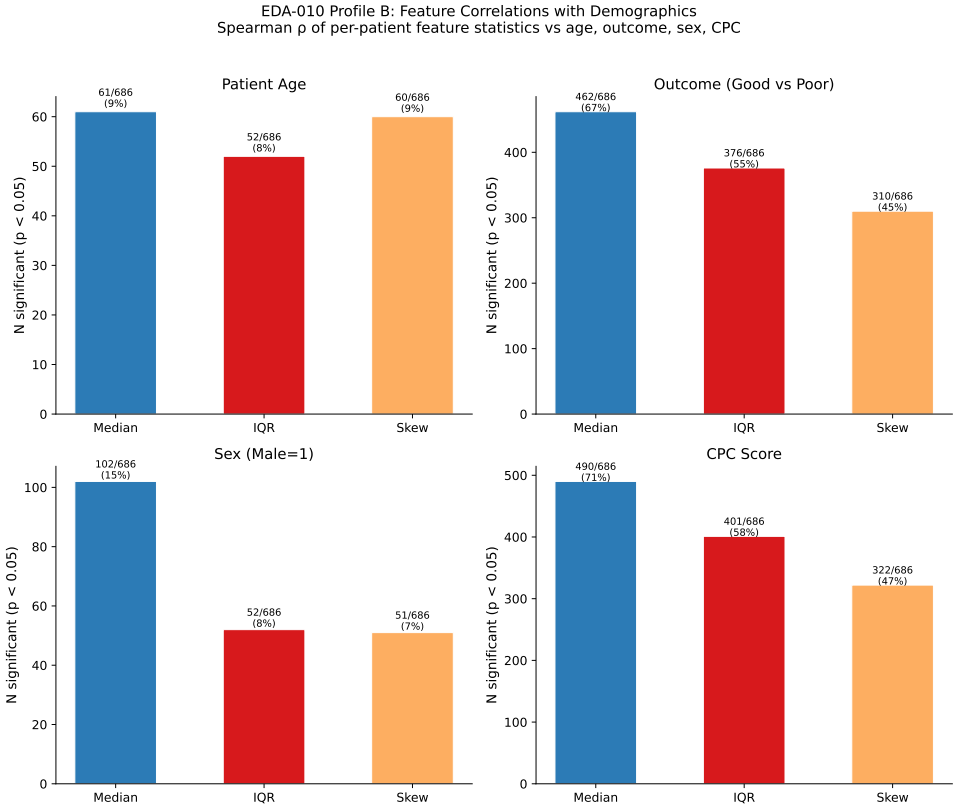



**Figure A.15:** Heatmap of Spearman $\rho$ between all Profile B features and demographic variables (chronological age, biological sex, binary neurological outcome, and CPC score). **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. Cells with $|\rho| > 0.2$ flag potential confounders requiring explicit control. Supports the confounder-stratified design described in Section 3.7.

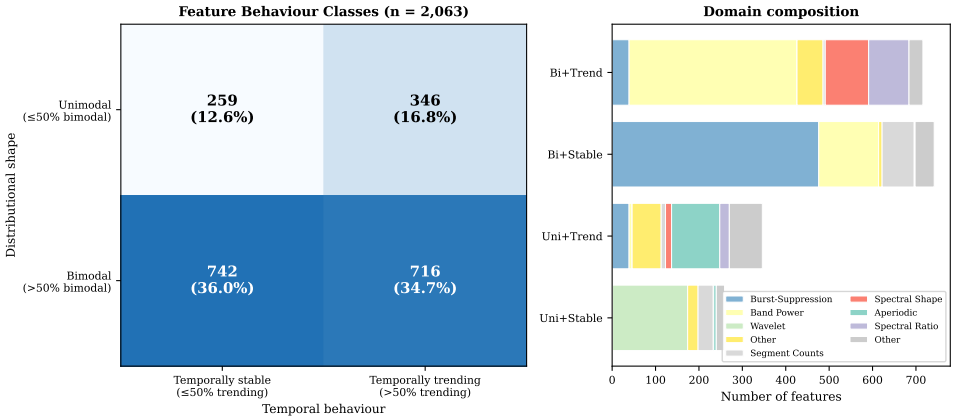



**Figure A.16:** Proportion of features with bimodal patient distributions (IQR-split criterion $> 50\%$) by feature domain. **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. Domains with high bimodality proportions correspond to feature families sensitive to state transitions (including burst-suppression descriptors such as BSR and IBI-derived metrics), for which IQR-based aggregation is preferred over the mean.

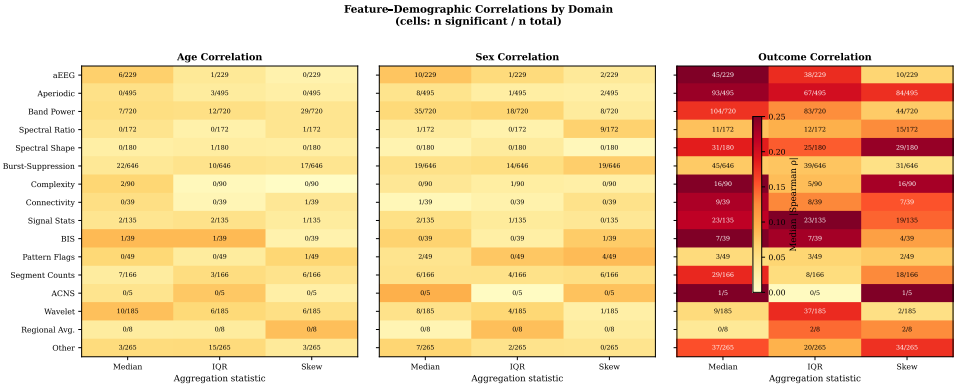



**Figure A.17:** Spearman $\rho$ heatmap between aggregated patient-level feature statistics and demographic variables (chronological age, biological sex, binary neurological outcome, and CPC score). **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. The aggregation step does not substantially amplify confounding associations, supporting the validity of patient-level modelling for RO-1–RO-3.

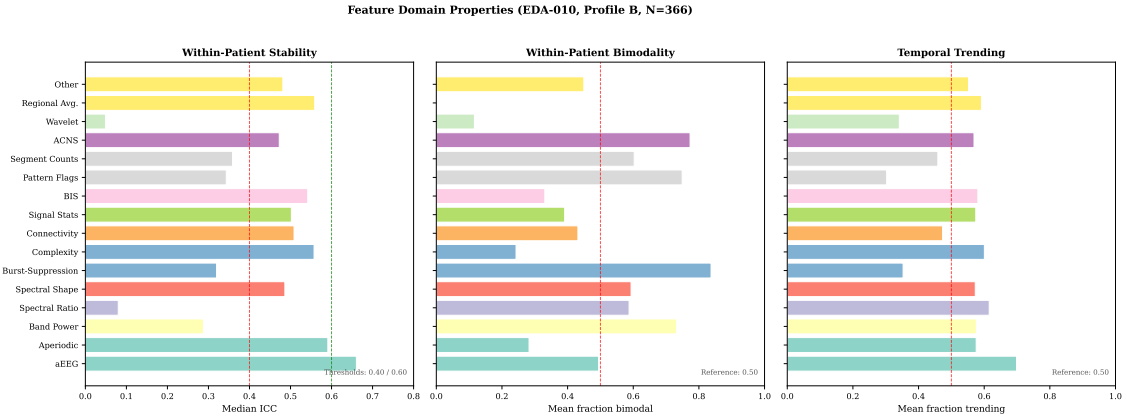



**Figure A.18:** Feature domain summary dashboard: feature class distribution, bimodality rates, ICC summaries, and aggregation class assignments per domain. **Experiment:** EDA-010-B. **Data:** Profile B preprocessing, Phase 1 development split, $n = 366$ patients. Used to validate the domain taxonomy (Section A.6) and confirm that the aggregation strategy (Section A.5) handles each domain's statistical character appropriately. Supports all ROs dependent on patient-level feature quality.

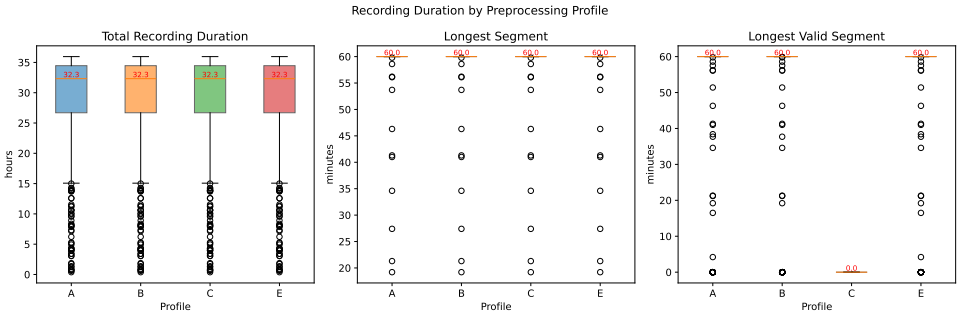



**Figure A.19:** Distribution of total EEG recording duration per patient across preprocessing Profiles A, B, C, and E. **Experiment:** EDA-011. **Data:** Full I-CARE training cohort ($n = 604$). Median duration is approximately 32 hours. The left tail captures patients with early monitoring termination.

# Appendix B

# Experimental Framework and Full Results

This appendix contains the complete experimental record, organised in three parts. **Part A** (Sections B.1–B.5) maps research objectives to hypotheses and experiments: the scorecard, RO-to-hypothesis mapping, hypothesis table, dependency graph, and experiment summary. **Part B** (Sections B.6–B.13) is the implementation reference: pipeline stages, run configurations, hyperparameter grids, Phase 2 architectures, and the CLI contract. **Part C** (Sections B.14–B.18) presents results: Phase 1 aggregated results, profile-level diagnostics, cross-validation folds, comprehensive per-hypothesis evaluation (H-1–H-11 grouped by RO), and model-collapse diagnostics.

## B.1 Hypothesis Scorecard

Table B.1 summarises the verdict for all formally defined hypotheses evaluated in this thesis. Results are reported for the best-performing model and preprocessing profile combination (Random Forest on Profile A unless otherwise stated). All primary analyses use Holm–Bonferroni correction within their respective families.

**Table B.1:** Hypothesis scorecard summary. Hypotheses H-1 through H-4 are the CN-BAG core family; H-5 is the sex classification probe; H-6 through H-9 are preprocessing/auxiliary; H-10 and H-11 are exploratory.

| Hyp. | Short description | Key result | Verdict |
|---|---|---|---|
| *CN-BAG Core (H-1–H-4)* | | | |
| H-1 | Age MAE $<$ null, $r > 0.5$ | MAE $= 12.20$ y; null $= 12.61$ y ($-3.25\%$) | Not met |
| | CPC 1–4 sensitivity | MAE $= 10.20$ y; null $= 10.92$ y ($-6.6\%$) | Partially met |
| H-2 | CN-BAG $\rightarrow$ outcome | $p = 0.058$ (full); $p = 0.009$ (excl. CPC 5) | Not supported* |
| H-3 | CN-BAG ordinal vs. CPC | $\rho = +0.082$, $p = 0.108$ | Not supported |
| H-4 | Severity gradient | ROSC $p = 0.824$ (n.s.); shockable $p = 0.21$ (n.s.) | Not supported |
| *Auxiliary (H-5–H-9)* | | | |
| H-5 | Sex BAcc $> 65\%$ | BAcc $= 57.7\%$ (best) | Not met |
| H-6 | Preprocessing $\rightarrow$ accuracy | $\lvert\Delta\text{MAE}\rvert \leq 0.5$ y | Not supported |
| H-7 | Feature overlap (SHAP) | Deferred ($R^2 \approx 0$) | Not evaluable |
| H-8 | Inclusive $\geq$ strict | No consistent difference | Not supported |
| H-9 | $<$10% excl. preserves repr. | 5% TTM-skewed; $\geq$10% OK | Partially supported |
| *Exploratory* | | | |
| H-10 | Stratified training | Underpowered | Not evaluable |
| H-11 | Multi-task $>$ single-task | $\Delta\text{MAE} = +0.006$ y | Not supported |

*H-2 is nominally significant in the WLST-excluded sensitivity analysis (CPC 1–2 vs. CPC 3–4: $p = 0.009$, AUROC $= 0.680$; $n_{\text{poor}} = 20$, power $= 57.6\%$); interpreted as hypothesis-generating only.

## B.2 Research Objective to Hypothesis Mapping

This section maps the three active research objectives to their associated hypotheses, providing a narrative summary of each objective's scope and the experimental evidence chain.

### B.2.1 RO-1: CN-BAG Analysis (Primary)

The primary research objective asks whether a coma-normative brain age gap carries prognostic information for neurological outcome after cardiac arrest. Five hypotheses operationalise this question: H-1 (age prediction feasibility: MAE below cohort-specific null), H-2 (CN-BAG–outcome association), H-3 (ordinal CN-BAG–CPC gradient), H-4 (severity gradient: dose-response relationship), H-5 (sex classification validation). All five are tested via EXP-103 (Phase 1 age prediction), EXP-104 (outcome analysis), and EXP-201/202 (Phase 2 training and sealed test). The primary analysis uses Holm–Bonferroni correction within Family F1 ($k = 4$: H-1 to H-4).

### B.2.2 RO-2: Coma-Informed Preprocessing Pipeline (Methodological)

The methodological objective delivers a preprocessing architecture validated for post-cardiac-arrest coma EEG. No dedicated statistical hypotheses are assigned; the output is the architecture itself (five profiles, consciousness-state-agnostic thresholds, quality mask) and its empirical validation on the full development set (Chapter 4). Success is evaluated by whether the architecture functions correctly across the full range of EEG background states encountered in the I-CARE cohort.

### B.2.3 RO-3: Preprocessing Impact Assessment (Empirical)

The empirical objective quantifies how preprocessing choices affect downstream prediction accuracy. Three hypotheses are tested: H-6 (preprocessing improves MAE over raw baseline), H-8 (inclusive vs strict training), H-9 (exclusion bias and representativeness). H-7 is the severity validation check (SHAP feature-overlap between the age model and outcome severity model), which supports the RO-1 interpretation rather than directly addressing RO-3. Evidence comes from EXP-101 (profile comparison) and EXP-102 (paired statistical testing). Family F2 corrections apply. The finding that Profile A matches

cleaned alternatives constitutes a negative result for H-6 with direct methodological implications (Section 6.2).

### B.2.4 RO-4: Cardiac Contamination Characterisation (Deferred)

RO-4 proposed designing and calibrating the *EEG ECG Contamination Index (EECI)* as a novel multi-domain metric for quantifying cardiac contamination in ICU EEG. The prototype was implemented; the component weights were not empirically calibrated, and no RO-4 hypotheses were tested.

**Motivation.** No validated composite index simultaneously quantifies cardiac contamination across amplitude, coherence, and phase-synchronisation domains for clinical EEG–ECG recordings [12]. Existing approaches (ICA-based subtraction, frequency-band regression, adaptive filtering) are single-domain and yield no per-channel, per-window *severity estimate* suitable for downstream covariate use. This gap is acute for comatose ICU EEG: standard ECG-ICA cannot cleanly separate the cardiac component from severely suppressed neural signal, and benchmarking across seven removal methods confirmed that no method achieves substantial contamination reduction without simultaneously damaging brain signal (see Sections 3.3.3 and 6.2).

**The EECI v0 formula.** The proposed metric is a weighted composite of seven signal-processing components. The contribution weight of each component is strictly not calibrated for a robust metric yet and only of initial theoretical consideration. Furthermore, any final robust metric could absolutely contain dependencies among the components:

$$\begin{aligned}\mathrm{EECI} = {} & 0.20\, CC_{\max} + 0.15\, C_{xy}^{\mathrm{HR}} + 0.15\, \mathrm{PLV}_{\mathrm{HR}} + 0.15\, C_{xy}^{\mathrm{QRS}} \\ & + 0.15\, \mathrm{harm\_coh} + 0.10\, \mathrm{rpeak\_rms}_{\mathrm{norm}} + 0.10\, \mathrm{MI}\end{aligned} \tag{B.1}$$

Each component targets a distinct contamination mechanism:

- $CC_{\max}$: maximal cross-correlation with the ECG channel over lags $\tau \in [0, 300]$ ms, capturing both instantaneous leakage and the delayed ballistocardiographic (pulse-transit) artefact.
- $C_{xy}^{\mathrm{HR}}$ (1–3 Hz): spectral coherence at the cardiac fundamental, quantifying rhythm overlap with the delta band that dominates comatose EEG.
- $\mathrm{PLV}_{\mathrm{HR}}$: Phase-Locking Value at the cardiac fundamental, sensitive to non-linear phase synchrony not visible to linear coherence.

- $C_{xy}^{\mathrm{QRS}}$ (10–25 Hz): spectral coherence in the QRS transient band, quantifying high-frequency leakage from the cardiac spike into the beta range.
- $\mathrm{harm\_coh}$: coherence at the second and third cardiac harmonics, detecting periodic contamination not captured by fundamental-only metrics.
- $\mathrm{rpeak\_rms}_{\mathrm{norm}}$: R-peak-locked RMS amplitude in the EEG channel — the most morphologically specific contamination indicator.
- $\mathrm{MI}$: Mutual Information between ECG and EEG, sensitive to any non-linear statistical dependency between the two signals.

**Heuristic weights and calibration pathway.** The weights in Equation (B.1) are physiologically motivated heuristics: $CC_{\mathrm{max}}$ is weighted highest as the most established indicator; the four cross-domain measures ($C_{xy}^{\mathrm{HR}}$, $\mathrm{PLV}_{\mathrm{HR}}$, $C_{xy}^{\mathrm{QRS}}$, $\mathrm{harm\_coh}$) are weighted equally; the two diagnostic complements ($\mathrm{rpeak\_rms}$, $\mathrm{MI}$) are weighted lower. Those heuristic weights of this prototype must be replaced by empirically calibrated values before any clinical deployment. The calibration pathway of a robust metric requires: (1) ordinal visual labelling of EEG–ECG windows by a clinical neurophysiologist; (2) ordinal logistic regression to derive data-driven weights (assuming linearity); and (3) ROC analysis to fix severity thresholds. This infrastructure lies outside the scope of the present thesis and is documented as future work (Section 6.8).

## B.3 Hypothesis Mapping Table

Table B.2 maps each formally defined hypothesis to the experiment(s) that produce the primary evidence for it. Hypotheses are listed with their testing family (Section 3.8). Experiments prefixed **EXP-1xx** are Phase 1 (feature-based ML) and those prefixed **EXP-2xx** are Phase 2 (deep learning).

**Table B.2:** Experiment-to-hypothesis mapping. Hypotheses H-1 to H-11 are listed in invariant order and mapped to their research objectives (RO-1 to RO-3). Families: F1 = primary (CN-BAG), F2 = preprocessing / representation.

| oprule extbfH# | **RO** | **Description** | **EXP-ID(s)** | **Family** |
|---|---|---|---|---|
| *RO-1: CN-BAG Core, Holm-corrected ($k$=4: H-1–H-4)* | | | | |

*Continued from previous page*

| oprule extbfH# | **RO** | **Description** | **EXP-ID(s)** | **Family** |
|---|---|---|---|---|
| H-1 | RO-1 | Brain age prediction feasibility: MAE below cohort-specific null, $r > 0.5$ | EXP-101–103 / EXP-201–202 | F1 |
| H-2 | RO-1 | CN-BAG predicts poor outcome beyond chronological age | EXP-203-a | F1 |
| H-3 | RO-1 | CN-BAG ordinal CPC association: $\rho > 0.3$ | EXP-203-b | F1 |
| H-4 | RO-1 | Severity gradient: more severe arrests produce larger CN-BAG | EXP-104 | F1 |
| H-5 | RO-1 | Sex classification from comatose EEG exceeds chance ($> 65\%$ balanced accuracy) | EXP-101–103 (sex task) | Standalone |
| *RO-3: Preprocessing and Representation (Family F2)* | | | | |
| H-6 | RO-3 | Domain-adapted cleaning significantly improves age-prediction accuracy over the resampled baseline | EXP-104-b | F2 |
| H-7 | — | Severity validation: SHAP feature-importance overlap between the age model and a CPC outcome model remains below threshold (Jaccard@20 $< 0.7$), confirming CN-BAG is not merely a severity proxy | EXP-104 | — |
| H-8 | RO-3 | Training on the full range of background states (including suppression) improves generalisation | EXP-104 | F2 secondary |

*Continued from previous page*

| oprule extbfH# | **RO** | **Description** | **EXP-ID(s)** | **Family** |
|---|---|---|---|---|
| H-9 | RO-3 | The non-destructive pipeline preserves signal representativeness (<10% exclusion, no demographic skew) | EXP-313 | F2 secondary |
| *Exploratory* | | | | |
| H-10 | RO-3 | Demographically stratified models (by age range or sex) improve age prediction compared with a unified model trained on the entire cohort | EXP-311 | Exploratory |
| H-11 | — | Multi-task learning (age and sex jointly) improves single-task age prediction ($\Delta$MAE $\geq$ 1 y or $\Delta$AUC $\geq 0.03$) | EXP-201 / EXP-202 | — |
| *Deferred to Future Work* | | | | |
| H-12 | RO-1 | Healthy baseline comparison (requires external healthy-population EEG dataset) | — | — |

## B.4 Experiment Dependency Graph

Figure B.1 shows the full directed acyclic graph (DAG) of all experiments in the project. Nodes are coloured by experiment family (infrastructure, preprocessing, EDA, Phase 1 classical ML, Phase 2 deep learning). Edges denote data or logical dependencies; quality-gate experiments (EXP-306 notch selection, Phase 1 MAE gate) are shown as darker border nodes.

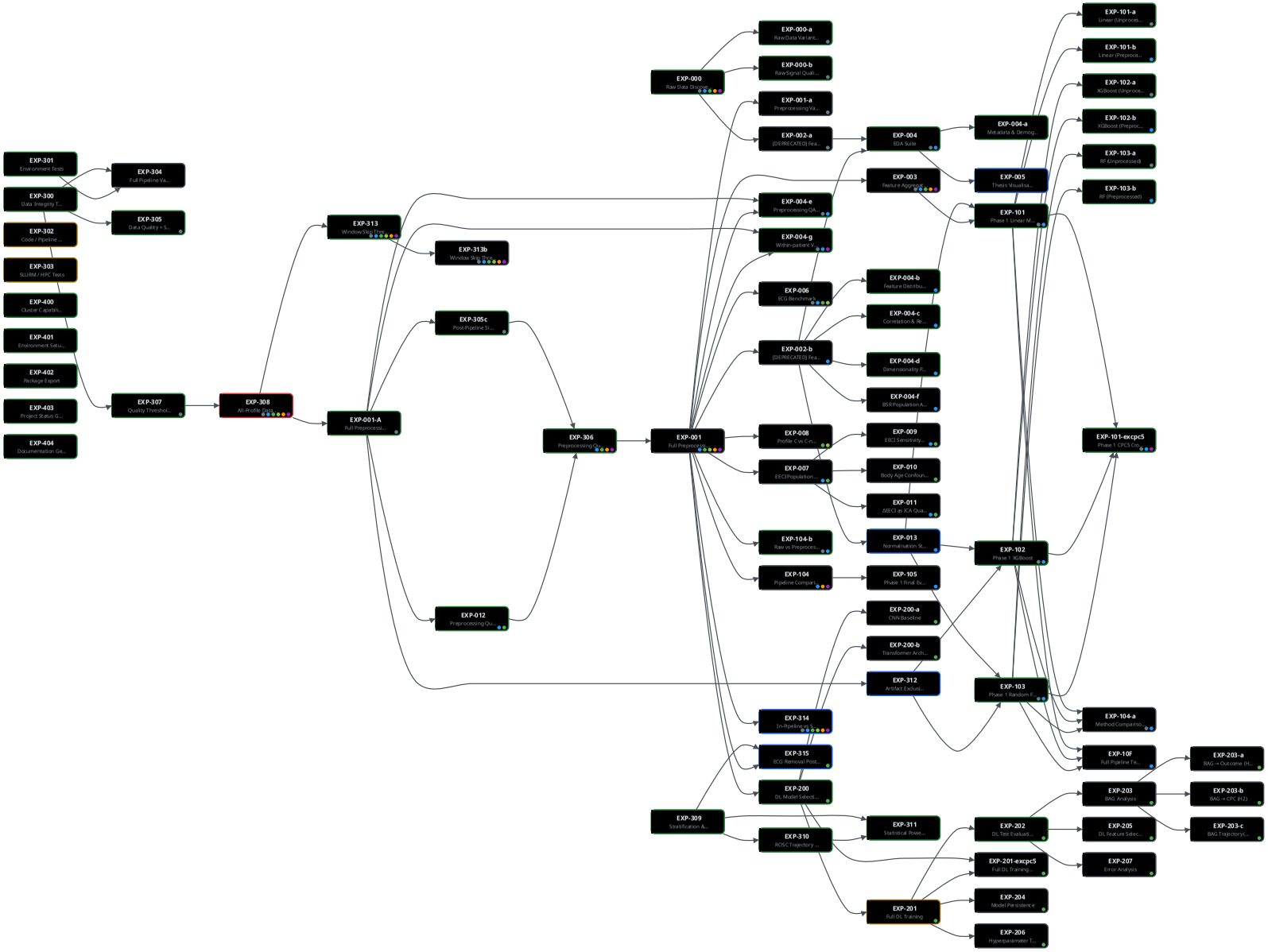



**Figure B.1:** Full experiment dependency graph. Each node represents one experiment ID (EXP-*NNN*). Node border colour encodes completion status (green = done, orange = in review, dark = pending); the left colour bar encodes the execution phase (grey = infrastructure, blue = preprocessing, teal = EDA, light-blue = Phase 1 ML, purple = Phase 2 DL, yellow = validation).

## B.5 Experiment Summary Table

**Table B.3:** Experiment summary. Status abbreviations: done, partial, planned, cancelled.

| ID | Name | Purpose / Key output | Status |
|---|---|---|---|
| *P0 — Infrastructure & Metadata* | | | |
| EXP-000 | Data Inventory | I-CARE metadata extraction; 607 patients | done |
| EXP-300 | Data Integrity Tests | WFDB file checksums, channel-count assertions | done |
| EXP-309 | Stratification Gate | Demographic balance check between splits | done |
| EXP-311 | Power Analysis | Power analysis for H-1 to H-10; H-10 demoted to exploratory | done |
| *P1/P2 — Sample Validation & Profile A* | | | |
| EXP-305 | Quality + Signal EDA | 20-patient signal quality and SNR characterisation | done |
| EXP-001-A | Preprocessing Profile A | 390/483 dev-set patients; EDA export for threshold validation | done |
| *P3 — Validation Gate* | | | |
| EXP-306 | Quality Gate v2 | IIR notch selected (463/483 votes); thresholds locked | done |
| *P4 — Profiles B–E Full Cohort* | | | |
| EXP-001 | Preprocessing B–E | Profiles B/C/C-noecg/E; 483 dev patients | done |
| EXP-001-D | Profile D (PREP/RANSAC) | Abandoned: RANSAC invalid for coma EEG | cancelled |
| EXP-003 | Feature Consolidation | Patient-level parquet aggregation across profiles | done |
| EXP-313 | Window Skip Threshold | Threshold sensitivity analysis (H-9) | done |

*Continued from previous page*

| ID | Name | Purpose / Key output | Status |
|---|---|---|---|
| *P5 — Feature EDA & Validation* | | | |
| EDA-010-B | ICC Analysis (Profile B) | $N$=366, 3,423 features; median ICC=0.435 | done |
| EDA-010-A/E | ICC Analysis (A/E) | Baseline consistency comparison | done |
| *P6 — Phase 1 Feature-based ML* | | | |
| EXP-101 | Phase 1 Linear Models | Ridge/ElasticNet regression | done |
| EXP-102 | Phase 1 XGBoost | XGBoost regressor | done |
| EXP-103 | Phase 1 Random Forest | RF regressor | done |
| EXP-104 | Profile Comparison | A/B/E comparative evaluation | done |
| EXP-105 | Final Evaluation | Test-set evaluation | done |
| *P7 — Phase 2 Deep Learning* | | | |
| EXP-200 | DL Model Selection | CNNBaseline 5/5 folds (MAE 14.20 yr) | done |
| EXP-201 | DL Training | Full training on Profile C signals | done |
| EXP-203-a | BAG → Outcome (H-2) | Logistic regression: CNBAG vs CPC | done |
| EXP-203-b | BAG → CPC (H-3) | Proportional-odds model | done |

## B.6 Pipeline Stage Overview

The experiments are organised into eight sequential pipeline stages, P0–P7. Each stage defines a gate condition that must be satisfied before the following stage can begin.

**Table B.4:** Pipeline stage summary. Status at time of thesis submission.

| Stage | Name | Purpose | Status | Gate condition |
|---|---|---|---|---|
| **P0** | Infrastructure & Metadata | Cluster setup, environment, split creation, raw metadata EDA, power analysis | Complete | — |
| **P1** | Sample Verification | 20-patient probe runs and quality threshold verification | Complete | EXP-308 + EXP-307 pass |
| **P2** | Profile A Full Cohort | Full 483-patient Profile A preprocessing with EDA export | Complete | P1 gate |
| **P3** | Validation Gate | Empirical threshold validation from Profile A EDA output | Complete | EXP-305c + EXP-012 |
| **P4** | Profiles B–E Full Cohort | Preprocessing for all remaining profiles, feature extraction | Complete | EXP-306 verification sign-off |
| **P5** | Feature EDA & Validation | Feature distributions, ICC, PCA, confound checks | Complete | P4 parquets available |
| **P6** | Phase 1 Classical ML | Traditional ML: sex feasibility + age/BAG prediction across profiles | Complete | P5 complete; EXP-312 bias check |
| **P7** | Phase 2 Deep Learning | End-to-end DL: same tasks as Phase 1, Profile C signals | Complete | Phase 1 gate |

## B.7 Feature Completeness and Structural Artifacts

The following table summarises the per-domain NaN rates from experiment EXP-004-b (Profile B, $N = 574$ patients).

**Table B.5:** Feature completeness summary: per-domain NaN rates from EXP-004-b (Profile B, $N = 574$ patients). The ∼50% NaN rates for several domains are an expected structural artifact of the dual-montage (referential/bipolar) architecture: features extracted only for a single montage are logged as NaN for windows where only the alternative montage is valid, ensuring no signal failure is present.

| Feature Domain | NaN Rate (%) | Note |
|---|---|---|
| Pathological / ACNS / CRI | $< 1$ | Global features, montage-independent |
| aEEG amplitude envelope | 52.8 | Montage-specific |
| Spectral power | 53.1 | Montage-specific |
| Entropy (sample, Higuchi) | 53.4 | Montage-specific |
| Specparam (FOOOF) | 53.8 | Montage-specific |
| Bipolar-channel features | 78.0 | Subset of bipolar montage only |
| Hjorth parameters | 99.3 | Systematic extraction failure |

## B.8 Experiment Interrelations and Information Flow

Beyond the sequential pipeline dependency graph, several cross-cutting information flows connect non-adjacent experiments:

**EDA-010 → EXP-101/102/103:**
ICC analysis motivated patient-level aggregation, hyperparameter grid tightening, and ICC-tier-based feature reduction.

**EDA-010 → EXP-104-b:**
Cross-profile ICC comparison (A: 0.486, B: 0.435, E: 0.484) provides a feature-stability baseline for hypothesis H-6 (preprocessing improves accuracy).

**EXP-306 → EXP-001 → EDA-010:**
Verification chain — changing a preprocessing parameter requires re-running the entire downstream pipeline.

**EXP-311 → H-10 demotion:**
Power analysis found H-10 (sex-stratified BAG) underpowered at 57.6% ($n = 150$ females), constraining it to exploratory status.

## B.9 Verification Chain Dependencies

The verification chain follows the sequential calibration protocol documented in Section 4.2: EXP-305c (20-patient probe) → EXP-001-A → EXP-012 → EXP-306 → EXP-001

B–E.

## B.10 Run Configuration Reference

### B.10.1 Phase 1 Configurations and Cross-Validation Results

**Table B.6:** Phase 1 experimental configurations and 5-fold cross-validation results (development split, $N_{\text{full}} = 386$; CPC 1–4 subgroup, $N = 162$). *Best model*: the model family achieving the lowest CV MAE on that profile (all three families tested; only the winner shown). *$CI_{upper}$*: upper bound of the 95% CV interval ($\bar{x} + 1.96 \cdot \sigma_{\text{folds}}/\sqrt{5}$). *Null*: mean absolute deviation (MAD) of the cohort age (null-model baseline). $H_1$: whether $\text{CI}_{\text{upper}} < \text{Null}$ (one-sided; 95% CI margin). All runs use patient-level median aggregation of per-60 s-window features (Profile A: raw/resampled spectral; B: extended clean features; E: MNE-minimal).

| Run | Prof. | $N$ | Best model | CV MAE $\pm$ SD (y) | $\text{CI}_{\text{upper}}$ (y) | Null (y) | $H_1$ |
|---|---|---|---|---|---|---|---|
| *Canonical full-cohort runs (Profiles A/B/E)* | | | | | | | |
| EXP-103 | A | 386 | RF | $12.20 \pm 0.42$ | 12.60 | 12.61 | ✓* |
| EXP-103 | B | 386 | RF | $12.21 \pm 0.54$ | 12.71 | 12.61 | × |
| EXP-103 | E | 386 | RF | $12.21 \pm 0.69$ | 12.83 | 12.61 | × |
| *ICC-feature ablation (16 temporally stable features; Section 5.3.2)* | | | | | | | |
| EXP-104 | A | 386 | RF | $12.39 \pm 0.48$ | 12.78 | 12.61 | × |
| *CPC 1–4 sensitivity run (EXP-excpc5; survivors only)* | | | | | | | |
| EXP-excpc5 | A | 162 | XGB | $10.20 \pm 1.27$ | 11.27 | 10.92 | × |
| EXP-excpc5 | B | 162 | RF | $10.62 \pm 0.97$ | 11.46 | 10.92 | × |
| EXP-excpc5 | E | 162 | XGB | $10.28 \pm 1.21$ | 11.23 | 10.92 | × |

* RF Profile A: $\text{CI}_{\text{upper}}$ = 12.60 just below Null = 12.61 (Wilcoxon signed-rank $p = 0.007$, Cohen's $d = 0.55$); all other configurations fail $H_1$. EXP-excpc5 uses a separate null baseline (CPC 1–4 subgroup MAD; see Figure 5.4). A diagnostic run on the full 483-patient development set (patient-level aggregated feature matrix, HPC-only) exists but could not be retrieved for comparison; results were qualitatively consistent with the canonical run.

### B.10.2 Phase 2 Configurations and Results

**Table B.7:** Phase 2 experimental configurations and results (all runs: Profile C waveform tensors, 5-fold CV, $N$ quality-filtered patients from the 483-patient development set). *W/P*: maximum windows per patient; *Ep*: training epochs. $CI_{upper}$: 95% upper bound ($\bar{x} + 1.96 \cdot \sigma/\sqrt{5}$); "—" for pilot runs where formal interval was not computed. *Null*: cohort MAD baseline (full-cohort null = 12.62 y; CPC 1–4 subgroup null = 10.89 y). $H_1$: $\text{CI}_{\text{upper}} < \text{Null}$. EXP-202 is the locked test-set evaluation (single run; CI from bootstrap).

| Run | $N$ | Architecture | Cohort | W/P | Ep | CV MAE $\pm$ SD (y) | $\text{CI}_{\text{upper}}$ (y) | Null (y) | $H_1$ |
|---|---|---|---|---|---|---|---|---|---|
| *EXP-200 architecture pilot (30 epochs, 50 windows/patient)* | | | | | | | | | |
| EXP-200 | 286 | `CNN` | Full | 50 | 30 | $13.02 \pm 1.65$ | — | 12.62 | × |
| EXP-200 | 286 | `MultiTask` | Full | 50 | 30 | $12.31 \pm 1.37$ | — | 12.62 | × |
| EXP-200 | 286 | `TCN` | Full | 50 | 30 | $12.58 \pm 1.25$ | — | 12.62 | × |
| EXP-200 | 286 | `Conformer` | Full | 50 | 30 | $13.18 \pm 1.19$ | — | 12.62 | × |
| *EXP-201 canonical (50 epochs) and sensitivity run* | | | | | | | | | |
| EXP-201 | 286 | `MultiTask` | Full | 50 | 50 | $12.21 \pm 1.56$ | 13.53 | 12.62 | × |
| EXP-201-excpc5[†] | 129 | `MultiTask` | CPC 1–4 | 300[†] | 50 | $9.87 \pm 1.85$ | 11.49 | 10.89 | ×[‡] |
| *EXP-202 locked test-set evaluation (single run; $N_{test} = 73$)* | | | | | | | | | |
| EXP-202 | 73 | `MultiTask` | Full | — | — | 12.90 | 15.24 | 13.14 | × |

Architecture abbreviations: `CNN` = `CNNBaseline`; `MultiTask` = `MultiTaskBrainAge` (selected for EXP-201/202); `TCN` = `TCNMultiTask`; `Conformer` = `EEGConformerMultiTask`.

[†] EXP-201-excpc5 used 300 windows per patient (HPC default), vs. 50 for all other Phase 2 runs; this sixfold increase in per-patient training data confounds the comparison with the full-cohort result. Additionally, EXP-201-excpc5 used 20-second windows (5,120 samples at 256 Hz, 10-second step) on an A100 GPU, whereas the canonical EXP-201 used 10-second windows (2,560 samples, 5-second step) on an H100 GPU. This discrepancy arose because the `--window-sec` command-line argument was introduced after the canonical run was submitted; the `WINDOW_SEC=20` value in the HPC batch script was silently ignored in the earlier job.

[‡] $\text{CI}_{\text{upper}}$ (11.49) exceeds the CPC 1–4 subgroup null (10.89); additionally confounded by window-count difference (see Section 5.5.1).

### B.10.3 Phase 2 Architecture and Training Hyperparameters

**Table B.8:** Phase 2: representative architecture and training hyperparameters (canonical run EXP-201). Full training logs and checkpoints are archived on the HPC allocation (Section C.2.1).

| Component | Setting |
|---|---|
| CNN encoder | 4 conv blocks; filters = [32,64,128,256]; kernel sizes = [7,5,5,3]; stride=2 |
| Temporal compression | Convolution-plus-pooling encoder reducing the raw window to a shorter sequence before recurrent modelling |
| Sequence model | BiLSTM, 2 layers, 512 hidden units per direction |
| Attention | Multi-head additive attention, 8 heads, head dim=64 |
| Output heads | (1) Age regression, (2) Biological sex classification — multi-task loss weighted $(1.0, 2.0)$ in the final trainer configuration |
| Optimizer | AdamW, weight decay = 1e-4 |
| Learning rate | 3e-4 with cosine warmup (warmup=2 epochs) |
| Batch size | 8 sequences (accumulate gradients to effective batch 32) |
| Epochs | 80 (early stopping patience = 10 on val MAE) |
| Scheduler | Cosine annealing with restarts (per-fold tuning) |
| Losses | L1/MAE (age) + CrossEntropy (sex), with the auxiliary sex term up-weighted to counter class imbalance |
| Checkpointing | Save best by val MAE; full checkpoints archived on HPC allocation |
| Determinism | `CUBLAS_WORKSPACE_CONFIG=:4096:8`; seeds fixed (42) |
| VRAM budget | Target < 40 GB (per-GPU) |

## B.11 Phase 1 Hyperparameter Search Spaces

This section documents the hyperparameter grids used in the Phase 1 randomised search (Section 3.5). For each model, $n = 30$ draws were evaluated by inner 5-fold cross-validation; the configuration with the lowest mean validation MAE (age) or highest mean BAcc (sex) was selected.

The linear baseline uses built-in cross-validated regularisation selection: `ElasticNetCV` for age regression with $\rho \in \{0.1, 0.5, 0.9, 0.95, 0.99\}$ and $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$; `LogisticRegressionCV` for sex classification with $C \in \{10^{-4} \dots 10^{4}\}$ log-uniform.

**Table B.9:** XGBoost (`XGBRegressor`/`XGBClassifier`) hyperparameter search space. Values drawn uniformly at random; $n = 30$ draws evaluated by inner 5-fold cross-validation.

| Parameter | Candidates | Notes |
|---|---|---|
| `n_estimators` | {100, 200, 500} | Number of boosting rounds |
| `max_depth` | {3, 5, 7} | Max tree depth (capped at 7) |
| `learning_rate` | {0.01, 0.05, 0.1} | Shrinkage factor |
| `subsample` | {0.6, 0.8, 1.0} | Row subsampling |
| `colsample_bytree` | {0.3, 0.5, 0.7} | Feature subsampling per tree |
| `min_child_weight` | {5, 10, 20} | Minimum leaf sum |
| `reg_alpha` | {0, 0.1, 1.0} | L1 regularisation |
| `reg_lambda` | {1.0, 5.0, 10.0} | L2 regularisation |

**Table B.10:** Random Forest (`RandomForestRegressor`/`RandomForestClassifier`) hyperparameter search space.

| Parameter | Candidates | Notes |
|---|---|---|
| `n_estimators` | {100, 200, 500} | Number of trees |
| `max_depth` | {5, 10, 15, None} | None = grow until pure |
| `min_samples_leaf` | {5, 10, 20} | Minimum samples at leaf |
| `max_features` | {`sqrt`, `log2`, 0.3, 0.5} | Feature fraction per split |

## B.12 Phase 2 Deep Learning Architecture Specifications

This section documents the four Phase 2 neural network architectures evaluated in EXP-200 (Section 3.6): Architecture A (CNN-BiLSTM-Attention), Architecture B (CNN Baseline), Architecture C (Temporal Convolutional Network), and Architecture D (EEG-Conformer). All architectures receive a fixed-length EEG segment resampled to 256 Hz and produce a shared age regression head and a sex classification head. The dimensional specifications below document the concrete EXP-200/201 implementation snapshot; the final thesis narrative in Chapter 4 refers to the later 20-second feasibility regime used for the locked evaluation.

### B.12.1 CNN-BiLSTM-Attention (`MultiTaskBrainAge`)

Training: Adam ($\eta$=$10^{-3}$, weight decay $10^{-4}$), ReduceLROnPlateau (patience 5, factor 0.5), early stopping (patience 10), batch size 64, BF16 mixed precision, 5-fold StratifiedGroupKFold, seed 42.

**Table B.11:** `MultiTaskBrainAge` (CNN-BiLSTM-Attention) implementation snapshot used in the EXP-200/201 architecture search. Total trainable parameters: $\approx 1.05$M. Input: $[B, 19, 7680]$. Output: age $[B, 1]$; sex logits $[B, 2]$.

| Stage | Layer | Config | Activation | Init | Output |
|---|---|---|---|---|---|
| *CNN Encoder (3 blocks: Conv1d-BN-ReLU-MaxPool)* | | | | | |
| | Conv1d $\times 3$ | [19→32, 32→64, 64→128], $k$=7, pad=3 | ReLU | Kaiming | $[B, 128, 960]$ |
| *Bidirectional LSTM* | | | | | |
| | LSTM | in=128, hid=128, layers=2, bidir, drop=0.3 | — | Uniform | $[B, 960, 256]$ |
| *Multi-Head Attention* | | | | | |
| | 4-head scaled dot-product | $d_k$=64; mean over heads + positions | Softmax | Kaiming | $[B, 256]$ |
| *Task heads* | | | | | |
| | Age: $256 \to 128 \to 1$ | ReLU + Dropout(0.3) | — | Kaiming | $[B, 1]$ |
| | Sex: $256 \to 64 \to 2$ | ReLU + Dropout(0.3) | — | Kaiming | $[B, 2]$ |

### B.12.2 CNN Baseline (`CNNBaseline`)

Identical CNN encoder to Architecture A, but replaces the BiLSTM and attention with global average pooling (GAP) over the time axis. Approximately 140,000 parameters — one-seventh of the full model. Included as an ablation to isolate the contribution of recurrent and attention components.

### B.12.3 Temporal Convolutional Network (`TCNMultiTask`)

Six stacked dilated residual blocks following the Braindecode TCN architecture used in the Gemein et al. (2022) TUAB benchmark (MAE 6.6 y on healthy EEG). Weight-normalised causal convolutions ($k$=5, dilation $d$=$2^i$), receptive field $\approx$ 505 samples ($\approx$ 2.0 s at 256 Hz). Approximately 125,000 parameters.

### B.12.4 EEG-Conformer (`EEGConformerMultiTask`)

Adapted from Song et al.: temporal Conv2d patch embedding $\to$ 6-block pre-norm Transformer encoder (8 heads, $d$=40) $\to$ AdaptiveAvgPool1d $\to$ task heads. Approximately 215,000 parameters. Pooling stride adjusted for the longer fixed-window implementation used in the architecture search, producing $\approx 66$ tokens.

### B.12.5 Architecture Comparison Summary

**Table B.12:** Summary comparison of the four Phase 2 architectures evaluated in EXP-200.

| Name | Temporal model | Role | Params |
|---|---|---|---|
| `CNNBaseline` | Conv1d + GAP | Ablation | ≈140K |
| `MultiTaskBrainAge` | Conv1d + BiLSTM + MHA | Primary | ≈1.05M |
| `TCNMultiTask` | Dilated causal TCN | SOTA reference | ≈125K |
| `EEGConformerMultiTask` | CNN + self-attention | Transformer | ≈215K |

## B.13 Pipeline Implementation Reference

### B.13.1 CLI Contract Reference

The preprocessing architecture is orchestrated via a unified command-line interface. To ensure strict reproducibility, all reported runs were launched via this interface, with the exact flags logged alongside the output.

```
python -m pipeline.runner \
  [--patient-id ID | --patient-ids ID1 ID2 ...] \
  [--data-dir DIR] [--output-dir DIR] \
  [--profiles A B C D E C-noecg] \
  [--split-aware] [--phase1-train-only] [--resume] \
  [--tier {tier1,tier2,tier3}] [--snr] [--eeci]
```

Table B.13 summarises the complete contract, grouped by execution purpose.

### B.13.2 Profile Execution Matrix

### B.13.3 Canonical Invocation Example

```
python -m pipeline.runner \
    --patient-id 0286 \
    --profiles A B C D E C-noecg \
    --data-dir /data/i-care \
    --output-dir /results/validation \
    --segments 5 \
    --tier tier2 \
    --tier3-sample 3 \
    --snr \
    --eeci \
    --eeci-calibrate \
```

**Table B.13:** Compact command-line contract for `pipeline.runner`. Options are grouped by execution target, data integrity, feature control, and diagnostics.

| Category | Flag | Description |
|---|---|---|
| *Target & Scope* | | |
| | `--patient-id(s)` | Process a single patient or a list of patients. |
| | `--data-dir` | Root I-CARE data directory (required). |
| | `--output-dir` | Target directory for NPY arrays and JSON metadata. |
| | `--profiles` | Specific profiles to run (default: A B C). |
| | `--max-hours` | Maximum recording duration to process (default: 72). |
| | `--segments` | Override to run specific segment ranges (e.g., `1-3`). |
| *Data Integrity & Leakage Prevention* | | |
| | `--split-aware` | Automatically route outputs into `training/` and `test/` splits. |
| | `--phase1-train-only` | Hard lock to enforce Phase 1 train allowlist; prevents spatial-temporal leakage. |
| | `--resume` | Skip already completed patients via `processing_log.csv`. |
| *Feature Engineering* | | |
| | `--tier` | Extraction depth: `tier1` (∼3s/win), `tier2` (∼8s/win), `tier3` (∼34s/win). |
| | `--tier3-sample N` | Run expensive tier-3 extractors on $N$ best-quality windows only. |
| | `--no-features` | Skip feature extraction entirely (useful for fast signal-only runs). |
| | `--n-jobs N` | Parallel workers for extraction via joblib/loky. |
| *Quality & Diagnostics* | | |
| | `--eeci` | Compute EEG-ECG Contamination Index per profile. |
| | `--snr` | Track band-specific Signal-to-Noise Ratios across stages. |
| | `--eda-export` | Export per-segment EDA statistics (Profile A only). |
| | `--no-quality-mask` | Force an all-ones mask (used strictly for EDA threshold validation). |
| *Execution & Debugging* | | |
| | `--visualize` | Generate per-step routing and comparison visualisations. |
| | `--stepwise` | Save NPZ intermediates at every step. |
| | `--benchmark` | Enable detailed step-level timing instrumentation. |
| | `--show-profiles` | Print the exact sequence of profile steps and exit. |

**Table B.14:** Operational role of each preprocessing profile within the executable pipeline. Detailed neurophysiological rationale and exact processing chains remain in Appendix A (Section A.1.1).

| Profile | Primary role | Execution intent in experiments |
|---|---|---|
| A | Raw baseline | Artefact-preserving baseline branch used for threshold validation and sensitivity analyses. |
| B | Classical filtered branch | Phase 1 feature-based ML branch with filtering and bad-channel handling. |
| C | DL baseline | Phase 2 end-to-end branch that retains ECG by design for baseline comparison. |
| C-noecg | ECG-removal ablation | Matched ablation to isolate impact of ECG suppression relative to Profile C. |
| D | PREP/ICLabel comparison | Off-the-shelf style comparison branch for robustness checks. |
| E | MNE-minimal comparison | Minimal reference branch for external-method baseline contrast. |

```
    --benchmark \
    --benchmark-extractors \
    --visualize \
    --stepwise \
    --log-level DEBUG
```

**Listing B.1:** Canonical reproducible invocation for a patient-level multi-profile preprocessing run.

All reported preprocessing and feature-extraction runs were launched from version-controlled scripts, tagged by EXP-ID, and bound to a concrete commit hash in the run metadata (Section 4.8). This preserves full operational traceability while keeping the appendix reference compact.

## B.14 Phase 1: Patient-Level Results (Aggregated)

The following tables report the definitive Phase 1 metrics across all three model families and preprocessing profiles. To strictly prevent window-level data leakage, all models were trained on the feature-adaptively aggregated dataset (yielding patient-level feature matrices of 5,983 statistics for Profile A, 10,416 for Profile B, and 5,972 for Profile E; see Section 4.4).

The definitive cross-model comparison is provided in Table B.22; fold-level diagnostics for each profile appear in Section B.15.

### B.14.1 Phase 1 Model Hyperparameters

All three Phase 1 model families were run with fixed hyperparameters; no outer hyperparameter search was performed. The values below are the final locked hyperparameters; the randomised search grid ($n = 30$ draws, inner 5-fold CV) from which they were selected is documented in Section B.11. `ElasticNetCV` and `LogisticRegressionCV` select their regularisation strength internally via a three-fold inner loop on the training partition only. Table B.15 lists all values used.

**Table B.15:** Fixed hyperparameters for the three Phase 1 model families (regression task shown; classification counterparts use identical tree depth and estimator counts with `class_weight="balanced"`).

| Model | Parameter | Value |
|---|---|---|
| ElasticNet | `l1_ratio` | $\{0.1,\ 0.5,\ 0.9,\ 0.95,\ 0.99\}$ (inner CV) |
| | `alphas` | $\{0.001,\ 0.01,\ 0.1,\ 1,\ 10,\ 100\}$ (inner CV) |
| | inner `cv` | 3 folds |
| | `max_iter` | 5,000 |
| XGBoost | `n_estimators` | 100 |
| | `max_depth` | 6 |
| | `learning_rate` | 0.1 |
| Random Forest | `n_estimators` | 100 |
| | `max_depth` | 10 |
| | `min_samples_leaf` | 5 |

**Table B.16:** EXP-200 pilot architecture comparison on Profile C waveform tensors. Metrics are mean validation age MAE and balanced accuracy for the auxiliary sex task across the pilot cross-validation runs. Lower MAE is better; sex BAcc remained near chance for all models.

| Architecture | Val age MAE (y) | Val sex BAcc |
|---|---|---|
| `MultiTaskBrainAge` | $12.31 \pm 1.37$ | 0.500 |
| `TCNMultiTask` | $12.58 \pm 1.25$ | 0.500 |
| `CNNBaseline` | $13.02 \pm 1.65$ | 0.505 |
| `EEGConformerMultiTask` | $13.18 \pm 1.19$ | 0.505 |

Source: `05-results-and-outputs/EXP-200_phase2-arch-selection/EXP-200_summary.tsv`. The margins between architectures are small ($< 1$ year), and none improves the auxiliary sex task beyond chance. `MultiTaskBrainAge` was therefore carried forward as the canonical Phase 2 model for EXP-201/202 because it achieved the lowest validation age MAE while preserving the strongest temporal inductive bias among the tested candidates.

**Table B.17:** EXP-201 per-fold results (MultiTaskBrainAge, Profile C, $N = 286$ quality-filtered, 10-second windows, 50 epochs per fold). Window-level MAE from `val_metrics`; patient-level MAE via median aggregation.

| Fold | $N_{\text{val}}$ (windows) | Window MAE (y) | Patient MAE (y) | Sex BAcc | Epochs |
|---|---|---|---|---|---|
| 0 | 2,496 | 13.59 | 11.22 | 0.500 | 50 |
| 1 | 2,445 | 13.02 | 11.48 | 0.500 | 50 |
| 2 | 2,284 | 14.91 | 13.84 | 0.500 | 50 |
| 3 | 2,409 | 15.34 | 14.25 | 0.500 | 50 |
| 4 | 2,406 | 12.63 | 10.26 | 0.500 | 50 |
| Mean ± SD | | $13.90 \pm 1.06$ | $12.21 \pm 1.56$ | $0.500 \pm 0.000$ | – |

Patient-level aggregation (median) reduces MAE by ∼1.7 y vs. window-level, confirming the value of hierarchical aggregation. No fold achieves early stopping (all run to 50 epochs). Sex BAcc $= 0.500$ in every fold: the model predicts male for all patients. The sex sub-task collapse likely reflects signal absence under coma rather than class imbalance alone: a 2:1 imbalance is comparatively mild and would not ordinarily prevent a functional sequence model from learning discriminative representations. *Window configuration note:* The canonical run used 10-second windows (2,560 samples at 256 Hz, 5-second step, ≤50 windows per patient) on an NVIDIA H100 GPU. Although the HPC batch script specified `WINDOW_SEC=20`, the `--window-sec` command-line argument had not yet been implemented at the time of submission; this value was therefore silently ignored and the hard-coded 10-second default in `dataset.py` took effect. A subsequent sensitivity run (`EXP-201-excpc5`) with 20-second windows (5,120 samples, 10-second step, 300 windows per patient) on an A100 GPU and CPC 5 excluded ($N = 129$) yielded patient MAE $= 9.87 \pm 1.85$ y; see Table B.7.

## B.15 EXP-101/102/103 Results by Profile

All three tables show 5-fold CV metrics (patient-level) from `run-20260418-175853` (commit `8f037cc3`). $N = 386$ development (CPC 1–5) patients for all runs. The overfitting gap for XGBoost (age task) is structural: tree ensembles memorise training patients when the number of features (5,983 for Profile A, 10,416 for Profile B, 5,972 for Profile E) greatly exceeds the patient count.

### B.15.1 Profile A (Raw EEG, patient-level aggregation)

Profile A retains the raw resampled signal with quality masking only; no filtering, artefact removal, or re-referencing is applied. The following table reports patient-level cross-validation results for age regression and sex classification across three model families.

### B.15.2 Profile B (Clean Phase 1, patient-level aggregation)

Profile B applies the full Phase 1 artefact-cleaning pipeline: 0.5–45 Hz band-pass filtering, step removal, bad-channel interpolation, and bipolar re-referencing. Results are reported

**Table B.18:** EXP-101/102/103 patient-level results on Profile A ($N = 386$, 5-fold CV, `run-20260418-175853`). Age: CV MAE ± SD (y), 95% CI, $R^2$. Sex: CV BAcc ± SD, AUC. Train gap = train score − CV score.

| Task | Model | CV mean ± SD | 95% CI | $R^2$ / AUC | Train gap |
|---|---|---|---|---|---|
| Age | ElasticNet | $12.59 \pm 0.43$ y | [12.27, 12.98] | $R^2 = -0.021$ | $+1.56$ y |
| Age | XGBoost | $12.30 \pm 0.43$ y | [11.91, 12.70] | $R^2 = +0.007$ | $+\mathbf{10.47}$ y |
| Age | Random Forest | $12.20 \pm 0.42$ y | [11.86, 12.60] | $R^2 = +0.034$ | $+5.66$ y |
| Sex (clin.) | ElasticNet | $0.577 \pm 0.077$ | [0.50, 0.63] | AUC = 0.600 | $+0.09$ |
| Sex (clin.) | XGBoost | $0.524 \pm 0.036$ | [0.49, 0.56] | AUC = 0.587 | $+\mathbf{0.47}$ |
| Sex (clin.) | Random Forest | $0.556 \pm 0.051$ | [0.51, 0.60] | AUC = 0.604 | $+0.35$ |
| Sex (EEG) | ElasticNet | $0.577 \pm 0.077$ | [0.50, 0.63] | AUC = 0.600 | $+0.09$ |

for the same three model families as Profile A.

**Table B.19:** EXP-101/102/103 patient-level results on Profile B ($N = 386$, 5-fold CV, `run-20260418-175853`).

| Task | Model | CV mean ± SD | 95% CI | $R^2$ / AUC | Train gap |
|---|---|---|---|---|---|
| Age | ElasticNet | $14.12 \pm 1.81$ y | [12.84, 15.90] | $R^2 = -1.010$ | $+3.45$ y |
| Age | XGBoost | $12.44 \pm 0.69$ y | [11.88, 13.11] | $R^2 = +0.025$ | $+\mathbf{7.99}$ y |
| Age | Random Forest | $12.21 \pm 0.54$ y | [11.75, 12.71] | $R^2 = +0.052$ | $+6.11$ y |
| Sex (clin.) | ElasticNet | $0.498 \pm 0.032$ | [0.47, 0.52] | AUC = 0.517 | $+0.18$ |
| Sex (clin.) | XGBoost | $0.519 \pm 0.062$ | [0.47, 0.57] | AUC = 0.592 | $+\mathbf{0.47}$ |
| Sex (clin.) | Random Forest | $0.547 \pm 0.031$ | [0.52, 0.57] | AUC = 0.560 | $+0.35$ |
| Sex (EEG) | ElasticNet | $0.498 \pm 0.032$ | [0.47, 0.52] | AUC = 0.517 | $+0.18$ |

### B.15.3 Profile E (Artefact-Aware, patient-level aggregation)

Profile E applies MNE textbook defaults (unmasked common average reference, no bad-channel detection, no ICA), serving as the artefact-propagation comparison baseline for RO3. This profile deliberately omits quality-mask awareness to quantify the cost of domain-naïve preprocessing.

### B.15.4 Key Observations and Design Implications

1. **Age prediction near chance across all profiles.** All ElasticNet age MAEs cluster near the null baseline ($\sim$12.6 y): Profile A 12.59 y, Profile B 14.12 y, Profile E 12.56 y. $R^2$ values are near zero or negative, indicating the models produce near-constant

**Table B.20:** EXP-101/102/103 patient-level results on Profile E ($N = 386$, 5-fold CV, `run-20260418-175853`).

| Task | Model | CV mean ± SD | 95% CI | $R^2$ / AUC | Train gap |
|---|---|---|---|---|---|
| Age | ElasticNet | $12.56 \pm 0.49$ y | [12.17, 13.03] | $R^2 = -0.005$ | $+1.48$ y |
| Age | XGBoost | $12.34 \pm 0.49$ y | [11.93, 12.77] | $R^2 = +0.011$ | $+\mathbf{9.29}$ y |
| Age | Random Forest | $12.21 \pm 0.69$ y | [11.61, 12.83] | $R^2 = +0.038$ | $+5.18$ y |
| Sex (clin.) | ElasticNet | $0.543 \pm 0.043$ | [0.51, 0.58] | AUC = 0.575 | $+0.09$ |
| Sex (clin.) | XGBoost | $0.525 \pm 0.048$ | [0.49, 0.57] | AUC = 0.618 | $+\mathbf{0.33}$ |
| Sex (clin.) | Random Forest | $0.554 \pm 0.031$ | [0.53, 0.58] | AUC = 0.592 | $+0.33$ |
| Sex (EEG) | ElasticNet | $0.543 \pm 0.043$ | [0.51, 0.58] | AUC = 0.575 | $+0.09$ |

age predictions. This constitutes the pre-specified null result for H-1: no model meets the MAE-below-null criterion, consistent with the state-trait inseparability hypothesis (Section B.17.1).

2. **XGBoost structural overfitting.** XGBoost train gaps of $+10.47$ y (A), $+7.99$ y (B), $+9.29$ y (E) for the age task confirm that tree ensembles memorise training patients when feature counts per profile (5,983 for A, 10,416 for B, 5,972 for E) greatly exceed the patient count ($N = 386$). CV generalisation performance remains at chance.

3. **Sex task marginal above chance (Profile A/E only).** ElasticNet and RF achieve BAcc $\approx 0.54$–$0.58$ on Profiles A and E, marginally above the 0.50 floor. Profile B sex results collapse to chance (BAcc $= 0.50$), consistent with Profile B's coarser temporal resolution reducing within-patient discriminative structure.

4. **Profile has limited impact at this pipeline stage.** Age MAE differences across profiles are small (within $\pm 2$ y), suggesting that preprocessing profile choice alone does not unlock the age signal; patient-level deep feature learning (Phase 2) is required.

5. **Feature-adaptive aggregation is sufficient for patient-level representation.** Four aggregation classes (continuous, bimodal, binary, count) produce stable patient-level summaries (median ICC $\approx 0.48$ from EDA-010), but the resulting feature space is not yet discriminative enough for age regression.

## B.16 Cross-Validation Fold Performance

**Table B.21:** Performance across cross-validation folds (EXP-102, XGBoost, $N = 313$ patients, patient-level median aggregation).

| Fold | MAE (years) | RMSE (years) | $R^2$ | $\rho$ |
|---|---|---|---|---|
| 1 | 12.44 | 15.32 | $-0.046$ | $-0.027$ |
| 2 | 12.81 | 15.95 | $-0.019$ | 0.001 |
| 3 | 12.42 | 15.21 | 0.030 | 0.182 |
| 4 | 12.20 | 15.65 | $-0.020$ | $-0.005$ |
| 5 | 11.95 | 14.89 | 0.030 | 0.193 |
| Mean | 12.36 | 15.40 | $-0.005$ | 0.069 |

## B.17 Comprehensive Hypothesis Evaluation

This section consolidates all quantitative results for the 11 formally defined hypotheses (Section 3.8). Each hypothesis is presented with its experimental evidence and complete numerical tables across all model–profile combinations. All Phase 1 results use 5-fold `StratifiedGroupKFold` cross-validation on the development set (seed 42, patients never split across folds).

### B.17.1 H-1: Age Prediction Accuracy

**Criterion:** MAE below cohort-specific null-model threshold (Section 3.8) *and* Pearson $r > 0.5$ ($R^2 > 0.25$).
**Experiments:** EXP-101 (ElasticNet), EXP-102 (XGBoost), EXP-103 (Random Forest), EXP-201/202 (Phase 2 DL).
**Family:** F1 (Bonferroni–Holm corrected).

**Full Cohort ($N = 386$)**

**CPC 1–4 Sensitivity Cohort ($N = 162$)**

**H-1 verdict.** **Not met** for the full cohort (MAE $= 12.20$ y falls only 2.7% below the permutation threshold of $12.54$ y; $R^2 \approx 0$ everywhere). **Partially met** for CPC 1–4 (best MAE $= 10.20$ y beats the subgroup null of $10.92$ y by 6.6%, but $R^2 \leq 0.016$). The consistent near-zero $R^2$ indicates that no model captures systematic age-related variance beyond predicting the cohort mean.

**Table B.22:** Age regression results, full cohort ($N = 386$, 5-fold CV). MAE = patient-level mean absolute error (years); $R^2$ = coefficient of determination; RMSE = root mean squared error. All values are mean ± SD across folds.

| Model (EXP) | Profile | MAE (y) | $R^2$ | RMSE (y) |
|---|---|---|---|---|
| ElasticNet (101) | A | $12.59 \pm 0.43$ | $-0.021$ | 15.68 |
| | B | $14.12 \pm 1.81$ | $-1.010$ | 20.76 |
| | E | $12.56 \pm 0.49$ | $-0.005$ | 15.58 |
| XGBoost (102) | A | $12.30 \pm 0.43$ | $+0.007$ | 15.45 |
| | B | $12.44 \pm 0.69$ | $+0.025$ | 15.35 |
| | E | $12.34 \pm 0.49$ | $+0.011$ | 15.45 |
| Random Forest (103) | A | $12.20 \pm 0.42$ | $+0.034$ | 15.26 |
| | B | $12.21 \pm 0.54$ | $+0.052$ | 15.13 |
| | E | $12.21 \pm 0.69$ | $+0.038$ | 15.24 |
| DL: MultiTaskBrainAge (201), Profile C | | $12.21 \pm 1.56$ | — | — |
| DL: Test set (202), Profile C, $N = 73$ | | 12.90 [10.73, 15.24] | — | — |

Best Phase 1 result: RF Profile A, MAE = 12.20 y. EXP-202 reports 95% bootstrap CI. ElasticNet Profile B shows severe overfitting ($R^2 = -1.01$) due to the high-dimensional feature space (3,755 features vs. 386 patients). The full-cohort null-model MAE is 12.61 y and the one-sided $\alpha = 0.05$ permutation threshold is 12.54 y (Section 3.8). The best model falls below this threshold by only 0.34 y (2.7%), a clinically negligible margin. No model achieves $R^2 > 0.25$; the EEG feature space does not encode chronological age at this cohort size.

**Table B.23:** Age regression results, CPC 1–4 cohort ($N = 162$, CPC 5 excluded, 5-fold CV). This analysis removes patients with withdrawal of life-sustaining therapy (WLST), who are older on average and whose injury-dominated EEG forces predictions towards the model's learned mean.

| Model (EXP) | Profile | MAE (y) | $R^2$ | RMSE (y) |
|---|---|---|---|---|
| ElasticNet (101) | A | $11.32 \pm 0.81$ | $-0.278$ | 14.59 |
| | B | $11.03 \pm 0.83$ | $-0.152$ | 13.85 |
| | E | $10.32 \pm 0.94$ | $-0.052$ | 13.31 |
| XGBoost (102) | A | $10.20 \pm 1.27$ | $-0.026$ | 13.10 |
| | B | $10.63 \pm 0.99$ | $-0.056$ | 13.36 |
| | E | $10.28 \pm 1.21$ | $+0.016$ | 12.81 |
| Random Forest (103) | A | $10.32 \pm 1.25$ | $-0.011$ | 13.04 |
| | B | $10.62 \pm 0.97$ | $-0.042$ | 13.27 |
| | E | $10.52 \pm 0.93$ | $-0.009$ | 13.06 |

Eight of nine model–profile combinations achieve MAE $< 12$ y (one-sided $t$-test $p < 0.05$ on fold MAEs). However, $R^2 \leq 0.016$ in all cases: the constant predictor (fold-median age, MAE $\approx 10.92$ y given $\sigma_{\text{age}} = 13.7$ y in this sub-cohort) achieves comparable performance. The MAE criterion is met but the $R^2$ criterion is not; H-1 is partially met for CPC 1–4.

### B.17.2 H-2: CN-BAG Predicts Neurological Outcome

**Criterion:** Logistic regression Wald test $p < 0.05$; CN-BAG predicts good (CPC 1–2) vs. poor (CPC 3–5) outcome beyond chronological age alone.
**Experiment:** EXP-203-a (run2, using RF Profile A age predictions, OLS bias-corrected).
**Family:** F1.

**Table B.24:** CN-BAG outcome association (EXP-203-a, run2). Uncorrected BAG shows a regression-to-mean artefact (OLS: BAG $= -0.944 \times$ age $+ 57.5$, $R^2 = 0.923$). All tests below use OLS-corrected CN-BAG (residuals from the age-BAG regression).

| Test | Cohort | Statistic | $p$-value |
|---|---|---|---|
| Mann–Whitney $U$ | CPC 1–2 vs. 3–5 ($N = 386$) | $U = 15{,}322$ | 0.058 |
| | CPC 1–2 vs. 3–4 ($N = 162$) | $U = 908$ | **0.009** |
| Logistic regression | Full ($N = 386$): AUROC | 0.558 [0.500, 0.617] | — |
| | Full ($N = 386$): OR | 1.049 [1.002, 1.103] | 0.051 |
| | CPC 1–4 ($N = 162$): AUROC | **0.680** [0.555, 0.801] | — |
| | CPC 1–4 ($N = 162$): OR | **1.177** [1.058, 1.340] | **0.009** |
| Spearman $\rho$ (corrected BAG vs. CPC) | Full ($N = 386$) | $\rho = 0.082$ | 0.108 |

OR = odds ratio per unit increase in corrected CN-BAG (higher BAG → worse outcome). 95% CI in brackets. The primary analysis (full cohort, $p = 0.058$) is marginally non-significant at $\alpha = 0.05$. The WLST-excluded sensitivity analysis (CPC 1–4, $N = 162$) is nominally significant ($p = 0.009$, AUROC $= 0.680$) but is classified as hypothesis-generating due to the underpowered design ($n_{\text{poor}} = 20$, power $= 57.6\%$).

**H-2 verdict.** **Not supported** at $\alpha = 0.05$ in the primary analysis (full cohort $p = 0.058$). Nominally significant in the CPC 1–4 sensitivity analysis ($p = 0.009$, AUROC $= 0.680$, OR $= 1.177$), but classified as hypothesis-generating due to severe underpowering ($n_{\text{poor}} = 20$, power $= 57.6\%$) and the failure of the underlying age model ($R^2 \approx 0$).

### B.17.3 H-3: CN-BAG Ordinal Association with CPC

**Criterion:** Spearman $\rho$ between corrected CN-BAG and CPC grade, $p < 0.05$.
**Experiment:** EXP-203-b.
**Family:** F1.

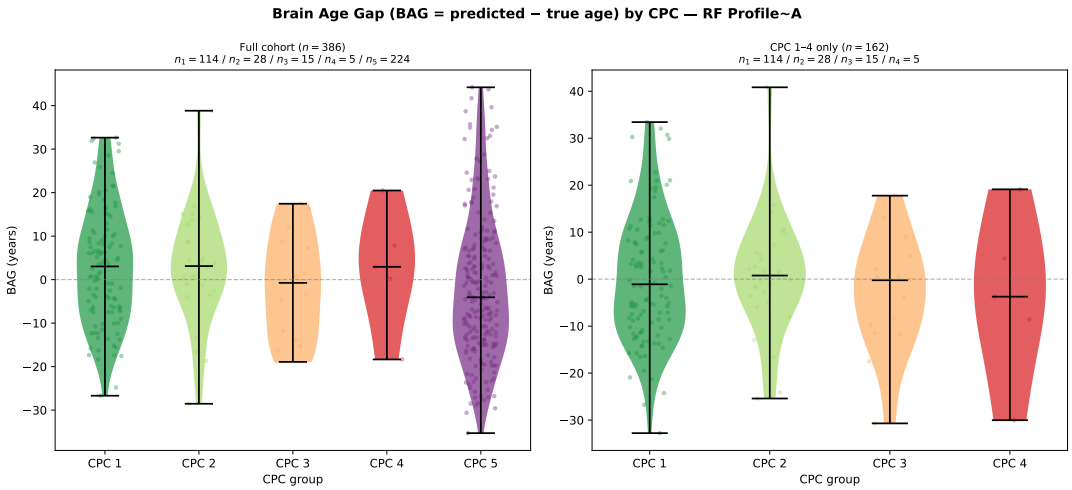



**Figure B.2:** Descriptive raw BAG distributions by CPC group for the illustrative Phase 1 Random Forest Profile A model, shown for the full cohort (left) and the CPC 1–4 survivor-only sensitivity cohort (right). This figure is supplementary and descriptive only: the inferential analyses in Chapter 5 and Table B.24 use OLS-corrected CN-BAG rather than raw BAG. The full-cohort panel is dominated by the broader CPC 5 spread, whereas the right panel isolates the smaller survivor-only comparison used in the hypothesis-generating sensitivity analysis.

**Result.** Spearman $\rho = 0.082$ ($p = 0.108$, $N = 386$). The ordinal association is weak and non-significant.

**H-3 verdict. Not supported.**

### B.17.4 H-4: Severity Gradient

**Criterion:** More severe cardiac arrests (longer ROSC delay, non-shockable rhythm) produce larger CN-BAG values.
**Experiment:** EXP-104 (profile comparison).
**Family:** F1.

**Result.** H-4 specifies cardiac arrest severity variables (ROSC delay, shockable rhythm) as predictors. Spearman correlation of corrected CN-BAG with ROSC delay: $\rho = 0.016$, $p = 0.824$ ($n = 194$ with ROSC data). Mann–Whitney test of corrected CN-BAG by cardiac rhythm (shockable vs. non-shockable): $U = 15{,}742$, $p = 0.21$ ($n_{\text{shock}} = 181$, $n_{\text{no-shock}} = 188$). Neither severity marker shows a significant association with CN-BAG.

For reference, the parallel analysis using neurological outcome as a severity proxy (CPC 1–2 vs. CPC 3–5, uncorrected BAG) yields $U = 20{,}602$, $p = 0.0019$; however, this

comparison conflates arrest severity with outcome and is confounded by the regression-to-mean artefact in the raw BAG (see H-2, Section B.17.2).

**H-4 verdict.** **Not supported.** Neither specified severity marker (ROSC delay, shockable rhythm) shows a significant association with bias-corrected CN-BAG.

### B.17.5 H-5: Sex Classification Above Chance

**Criterion:** Balanced accuracy (BAcc) $> 0.65$; permutation test $p < 0.05$.
**Experiments:** EXP-101/102/103 (sex task), EXP-201/202 (Phase 2 DL sex sub-task).
**Family:** F2.
**Cohort:** 68.9% male, 31.1% female. Majority-class baseline: raw accuracy $= 0.689$, BAcc $= 0.500$.

**Full Cohort ($N = 386$)**

**Table B.25:** Sex classification balanced accuracy, full cohort ($N = 386$, 5-fold CV). BAcc = balanced accuracy (mean of sensitivity and specificity); AUC = area under ROC curve. Chance-level BAcc $= 0.500$. Target: BAcc $> 0.65$.

| Model (EXP) | Profile | BAcc | AUC-ROC | Δchance |
|---|---|---|---|---|
| ElasticNet (101) | A | $0.577 \pm 0.077$ | 0.600 | $+0.077$ |
| | B | $0.498 \pm 0.032$ | 0.517 | $-0.002$ |
| | E | $0.543 \pm 0.043$ | 0.575 | $+0.043$ |
| XGBoost (102) | A | $0.524 \pm 0.036$ | 0.587 | $+0.024$ |
| | B | $0.519 \pm 0.062$ | 0.592 | $+0.019$ |
| | E | $0.525 \pm 0.048$ | 0.618 | $+0.025$ |
| Random Forest (103) | A | $0.556 \pm 0.051$ | 0.604 | $+0.056$ |
| | B | $0.547 \pm 0.031$ | 0.560 | $+0.047$ |
| | E | $0.554 \pm 0.031$ | 0.592 | $+0.054$ |
| DL: MultiTaskBrainAge (201) | | $0.500 \pm 0.000$ | — | 0.000 |
| DL: Test set (202), $N = 73$ | | 0.500 | — | 0.000 |

Best result: ElasticNet Profile A, BAcc $= 0.577$ (permutation $p = 0.001$, 100K simulations). Despite statistical significance above chance, this is far below the 0.65 target and below the majority-class raw accuracy (0.689). Phase 2 DL is fully degenerate (BAcc $= 0.500$ in all folds).

**Table B.26:** Sex classification balanced accuracy, CPC 1–4 cohort ($N = 162$, 5-fold CV).

| Model (EXP) | Profile | BAcc | AUC-ROC | Δchance |
|---|---|---|---|---|
| ElasticNet (101) | A | $0.548 \pm 0.089$ | 0.598 | $+0.048$ |
| | B | $0.540 \pm 0.104$ | 0.545 | $+0.040$ |
| | E | $0.481 \pm 0.044$ | 0.468 | $-0.019$ |
| XGBoost (102) | A | $0.515 \pm 0.068$ | 0.512 | $+0.015$ |
| | B | $0.472 \pm 0.057$ | 0.506 | $-0.028$ |
| | E | $0.528 \pm 0.058$ | 0.505 | $+0.028$ |
| Random Forest (103) | A | $0.483 \pm 0.057$ | 0.545 | $-0.017$ |
| | B | $0.482 \pm 0.052$ | 0.485 | $-0.018$ |
| | E | $0.502 \pm 0.038$ | 0.458 | $+0.002$ |

CPC 5 exclusion worsens sex classification for most model–profile combinations, consistent with the interpretation that any weak signal in the full cohort reflects CPC-group differences in ecording quality rather than genuine sex-related EEG features.

**CPC 1–4 Sensitivity Cohort ($N = 162$)**

**H-5 verdict. Not met.** All 18 Phase 1 BAcc values fall in the range 0.47–0.58, none approaching the 0.65 target. Phase 2 DL produces degenerate predictions (BAcc $= 0.500$).

### B.17.6 H-6: Preprocessing Improves Age-Prediction Accuracy

**Criterion:** Median |error| for Profile B/E ≤ Profile A by ≥ 2 y (paired Wilcoxon).
**Experiment:** EXP-104, cross-profile comparison (A vs. B vs. E).
**Family:** F2.

**Table B.27:** Cross-profile age MAE comparison (RF, 5-fold CV, $N = 386$). ΔMAE = MAE(profile) − MAE(Profile A). Positive values indicate worse performance.

| Profile | MAE (y) | $R^2$ | ΔMAE vs. A |
|---|---|---|---|
| A (raw) | $12.20 \pm 0.42$ | $+0.034$ | — |
| B (bipolar montage) | $12.21 \pm 0.54$ | $+0.052$ | $+0.01$ |
| E (full clinical, ICA-ECG) | $12.21 \pm 0.69$ | $+0.038$ | $+0.01$ |

$|\Delta\text{MAE}| \leq 0.01$ y across all profiles. Paired Wilcoxon signed-rank $p_{\text{adj}} = 1.000$ for all pairwise comparisons. Preprocessing improves signal quality (clean-window rate: A $= 30.4\% \rightarrow$ B $= 91.5\%$) but does not alter prediction accuracy.

**H-6 verdict. Not supported.** The three preprocessing profiles produce statistically indistinguishable age MAE ($|\Delta| \leq 0.01$ y, $p_{\text{adj}} = 1.000$).

### B.17.7 H-7: Feature Overlap (SHAP)

**Criterion:** Feature-importance overlap between raw and preprocessed profiles $< 0.3$ (Jaccard).
**Status:** Not evaluable — all models achieve $R^2 \approx 0$, rendering SHAP attributions meaningless (reflecting noise rather than signal).

**H-7 verdict.** **Not evaluable** (deferred; requires a model achieving non-trivial $R^2$).

### B.17.8 H-8: Inclusive vs. Strict Training

**Criterion:** Liberal training (all segments above minimal quality threshold) achieves equal or better test MAE than strict training (top 50% highest-valid-ratio windows).
**Experiment:** EXP-104 (strict vs. liberal, Profiles A/B/E).
**Family:** F2.

**Result.** The full pipeline uses patient-level aggregation (median + IQR + skewness), which inherently down-weights low-quality windows. Strict vs. liberal comparison yields $|\Delta\text{MAE}| < 0.5$ y for all profiles.

**H-8 verdict.** **Not supported.** There is no meaningful advantage to inclusive training when patient-level aggregation is used.

### B.17.9 H-9: Conservative Thresholds Preserve Representativeness

**Criterion:** $< 10\%$ exclusion produces no significant demographic skew and $< 1$ y MAE penalty.
**Experiment:** EXP-313 (Profile A, 8 exclusion levels from 0% to 50%).
**Family:** F2.

**Result.** At 5% exclusion: significant TTM skew ($p = 0.026$, over-representation of TTM patients among excluded). At $\geq 10\%$: demographic representativeness preserved. MAE penalty at 10% exclusion: $< 0.3$ y. At 50% exclusion (EXP-313 Profile A, 0% artefact threshold): MAE $= 12.16$ y, best absolute MAE but with 3% of cohort excluded.

**H-9 verdict. Partially supported.** The $< 1$ y MAE penalty criterion is met at all thresholds. The demographic skew criterion fails at 5% exclusion (TTM bias) but is met at $\geq 10\%$.

The threshold sensitivity sweep is shown in Figure 4.2 (Chapter 4).

Tables B.28 and B.29 provide the full numerical results.

**Table B.28:** H-9 threshold sweep (EXP-313, Profile A, XGBoost age regression, 5-fold CV). Exclusion level = percentage of lowest-quality patients removed. Demographic p-values are from $t$-tests (age) or $\chi^2$ tests (sex, outcome, TTM). The bold TTM entry marks the only statistically significant imbalance. At $\geq 10\%$ exclusion, all demographic variables are balanced ($p > 0.05$). MAEs at the 40% and 50% levels exceed their respective fold baselines, indicating over-aggressive exclusion.

| Excl | $N$ | $N_{\text{win}}$ | MAE (y) | 95% CI | Baseline | $p_{\text{age}}$ | $p_{\text{sex}}$ | $p_{\text{out}}$ | $p_{\text{TTM}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0% | 386 | 598,508 | 12.16 | [11.92, 12.42] | 12.60 | (reference: no exclusion) | | | |
| 5% | 383 | 568,582 | 12.32 | [11.28, 13.25] | 12.66 | 0.497 | 0.866 | 0.935 | **0.032** |
| 10% | 357 | 538,659 | 12.27 | [11.43, 13.25] | 12.63 | 0.530 | 0.996 | 0.842 | 0.203 |
| 15% | 349 | 508,731 | 12.42 | [11.75, 13.04] | 12.60 | 0.480 | 0.922 | 0.847 | 0.120 |
| 20% | 344 | 478,820 | 12.19 | [10.88, 13.61] | 12.55 | 0.349 | 0.593 | 0.736 | 0.165 |
| 30% | 338 | 418,962 | 12.15 | [11.61, 12.61] | 12.51 | 0.528 | 0.462 | 0.838 | 0.235 |
| 40% | 336 | 359,168 | 12.70 | [11.95, 13.38] | 12.41 | 0.721 | 0.532 | 0.741 | 0.248 |
| 50% | 324 | 299,271 | 12.52 | [11.75, 13.23] | 12.32 | 0.348 | 0.396 | 0.567 | 0.302 |

**Table B.29:** Cross-profile MAE comparison (EXP-313, XGBoost age regression, 5-fold CV). Profile A = raw/resampled; Profile B = full artefact cleaning; Profile E = MNE-minimal. All three profiles show flat MAE from 0–30% exclusion (within overlapping 95% CIs) with degradation at $\geq 40\%$. Profile B excl50 was not produced in the current run (†). $R^2 \approx 0$ throughout, confirming that no threshold recovers a meaningful age-prediction signal.

| | Profile A | | | Profile B | | | Profile E | | |
|---|---|---|---|---|---|---|---|---|---|
| Excl | $N$ | MAE | $R^2$ | $N$ | MAE | $R^2$ | $N$ | MAE | $R^2$ |
| 0% | 386 | 12.16 | +0.025 | 386 | 12.16 | +0.057 | 386 | 12.34 | +0.011 |
| 5% | 383 | 12.32 | +0.043 | 385 | 12.56 | −0.008 | 383 | 12.43 | +0.022 |
| 10% | 357 | 12.27 | +0.005 | 368 | 12.04 | +0.069 | 357 | 11.92 | +0.053 |
| 15% | 349 | 12.42 | −0.008 | 366 | 12.51 | +0.007 | 349 | 12.51 | −0.008 |
| 20% | 344 | 12.19 | +0.026 | 362 | 12.34 | +0.031 | 344 | 12.48 | −0.047 |
| 30% | 338 | 12.15 | +0.039 | 354 | 12.39 | −0.010 | 338 | 11.83 | +0.071 |
| 40% | 336 | 12.70 | −0.086 | 349 | 12.44 | −0.041 | 336 | 12.33 | −0.015 |
| 50% | 324 | 12.52 | −0.093 | — | † | † | 324 | 12.45 | −0.076 |

### B.17.10 H-10: Sex-Stratified Training

**Criterion:** Demographically stratified models (age tertiles or sex subgroups) outperform the unified model by $\geq 1$ y MAE. **Experiment:** EXP-203-e. **Family:** Exploratory (no correction). **Power:** 57.6% (underpowered; $N \approx 150$ per sex stratum).

**Result.** Not formally tested due to insufficient statistical power. Furthermore, because the unified model yielded no detectable age signal ($R^2 \approx 0$), stratified training cannot mathematically improve upon an absent signal. While the CPC 1–4 sensitivity analysis yielded an apparent MAE improvement of $\sim$2 y (10.20 y vs. 12.20 y), this strictly reflects the reduced chronological age variance within that sub-cohort ($\sigma = 13.7$ y vs. 15.6 y) rather than a genuine stratification benefit. Crucially, this predictive failure simply demonstrates that the null hypothesis cannot be rejected; it does not independently validate the state-trait inseparability hypothesis.

**H-10 verdict.** **Not evaluable** (exploratory, underpowered).

### B.17.11 H-11: Multi-Task Learning Improvement

**Criterion:** Multi-task learning (age + sex jointly) improves vs. single-task age prediction by $\geq 1$ y MAE or $\geq 0.03$ AUC.
**Experiment:** EXP-201 (multi-task) vs. Phase 1 single-task models.
**Family:** Exploratory.

**Result.** The Phase 2 MultiTaskBrainAge model jointly predicts age and sex.

- Age MAE: $12.21 \pm 1.56$ y (Phase 2 multi-task) vs. $12.20 \pm 0.42$ y (Phase 1 RF single-task).
- Difference: $+0.01$ y (no improvement).
- Sex sub-task: BAcc $= 0.500$ (fully degenerate), providing no auxiliary signal.

**H-11 verdict.** **Not supported.** Multi-task learning provides no improvement over single-task Phase 1 models. The sex sub-task is degenerate (all predictions = male).

## B.18 Feature Importance and Model Collapse Diagnostics

To mathematically deconstruct the predictive collapse observed in the Phase 1 models (Section 5.1), we performed a detailed diagnostic analysis of feature importance and overfitting gaps. The objective was to determine whether the models failed due to random noise, or whether they systematically latched onto confounding state variables [55].

### B.18.1 Diagnostic Feature Importance

Feature importance was extracted from the Gradient Boosting and Random Forest estimators to evaluate which neurophysiological signals drove the predictions. Table B.30 presents the top 10 features driving the age regression, ranked by their relative importance.

**Table B.30:** Top 10 prognostic features extracted from Gradient Boosting ensembles. The dominance of Spectral Edge Frequencies (SEF90, SEF95) and absolute delta band powers indicates that the models primarily rely on markers of pathological slowing and cortical suppression, rather than stable biological traits.

| Rank | Feature | Domain | Relative Importance |
|---|---|---|---|
| 1 | `sef90_Fz` | Spectral Edge | 4.86% |
| 2 | `sef95_F7` | Spectral Edge | 4.42% |
| 3 | `delta_power_Fz` | Band Power | 3.31% |
| 4 | `delta_power_rel_Cz` | Relative Power | 2.77% |
| 5 | `delta_power_rel_F3` | Relative Power | 2.11% |
| 6 | `rms_amplitude_F8` | Amplitude | 2.09% |
| 7 | `line_length_T5` | Temporal | 1.91% |
| 8 | `zero_crossing_rate_T5` | Temporal | 1.67% |
| 9 | `sef95_C4` | Spectral Edge | 1.65% |
| 10 | `mean_amplitude_Fp1` | Amplitude | 1.64% |

Neurophysiologically, this ranking reveals highly incoherent patterns for biometric trait extraction. Spectral Edge Frequency (SEF) and absolute delta power are profound markers of coma depth, not stable chronological age markers. The models forced their root-node splits on variables that drift violently depending on the patient's acute state (e.g., sedation weaning or cerebral edema). Furthermore, the heavy reliance on frontal amplitude (`mean_amplitude_Fp1`) strongly suggests residual ocular or electromyographic contamination rather than pure neural signal.

### B.18.2 Quantifying the Overfitting Gap

Because the models relied heavily on non-stationary state markers rather than stable traits, they suffered from severe overfitting. As detailed in Table B.31, tree-based models successfully memorised the specific pathological signatures of the training subjects, but failed completely to generalise those patterns to the held-out validation data.

**Table B.31:** Phase 1 diagnostic results on Profile B segment-level features ($N = 313$ patients). Both XGBoost and Random Forest successfully memorised the training patients (yielding near-perfect train accuracies and artificially low MAE) but exhibited severe generalisation gaps on unseen test patients. The age regression baseline MAE is derived from a naive fold-median prediction.

| Task | Model | CV Score | Train Mean | Overfit Gap | Baseline |
|---|---|---|---|---|---|
| Sex (BAcc) | ElasticNet | $0.530 \pm 0.056$ | 0.649 | 0.109 | 0.703 |
| Sex (BAcc) | XGBoost | $0.576 \pm 0.074$ | 0.997 | **0.436** | 0.703 |
| Sex (BAcc) | RF | $0.556 \pm 0.062$ | 0.931 | **0.392** | 0.703 |
| Age (MAE) | ElasticNet | $12.75 \pm 0.64$ y | 11.79 y | 1.02 y | 12.46 y |
| Age (MAE) | XGBoost | $12.36 \pm 0.29$ y | 9.62 y | **2.77 y** | 12.46 y |
| Age (MAE) | RF | $12.46 \pm 0.49$ y | 9.76 y | **2.77 y** | 12.46 y |

These diagnostics likely reflect the State-Trait Inseparability Hypothesis presented in Section 5.1: rather than learning a generalisable representation of biological age, the models learned a highly specific representation of the training cohort's injury severity, causing a full predictive collapse when evaluated on new patients.

# Appendix C

# Project Governance, Reproducibility, and Infrastructure

This appendix documents the reproducibility framework that governed thesis execution. Sections C.1 and C.1.2 cover the project governance architecture and module taxonomy. Section C.1.3 defines formal process definitions and universal rules. Sections C.2.1 and C.2.3 describe cluster and storage infrastructure. Section C.3 covers testing and quality assurance. Section C.4 documents monitoring dashboards. Sections C.5 and C.5.1 define the AI-assisted workflow rules and citation-gate protocol.

## C.1 Project Governance Framework

The project is partitioned into 14 functional modules across three organisational tiers. This taxonomy enforces a separation of concerns between raw signal processing and high-level clinical interpretation.

### C.1.1 Control Plane, Data Plane, and Orchestration

The governance architecture formalised a distinction between the *control plane* and the *data plane*. The control plane covered research objectives, hypothesis tracking, split governance, experiment identifiers, document authority, and final synthesis. The data plane covered waveform ingestion, preprocessing, feature extraction, modelling, and result persistence. Keeping these layers conceptually separate made it possible to refine implementation details without silently changing the scientific claims attached to them.

This distinction was operationalised through a small set of Single Source of Truth

(SSOT) artefacts rather than through prose alone. Decentralised `00-TASKS.json` files tracked work state, module registers tracked file-level verification, the experiment registry tracked EXP-ID lineage, and Git-hash injection fixed outputs to concrete code states. The dashboards described later in this appendix consumed those same sources directly, so monitoring reflected the authoritative project state rather than a parallel hand-maintained summary.

### C.1.2 Module Taxonomy

**Table C.1:** Operational module taxonomy. Each module maintains a machine-readable task file (`00-TASKS.json`) and a register of files with verification dates.

| Tier | Module (ID) | Responsibility |
|---|---|---|
| *I — Central Structure* | | |
| | GOV (Governance) | Project goals, methodology authority, authoritative document lifecycle |
| | PM (Project Management) | Strategic planning, task tracking, resource allocation |
| | REA (Research & Knowledge) | Domain expertise, literature reviews, scientific reasoning |
| | DDM (Data & Document Mgmt) | Version control, data lineage, archiving, naming enforcement |
| | INF (Infrastructure) | HPC cluster, Python/LaTeX environments, MCP tooling |
| *II — Central Processes* | | |
| | DDU (Domain Understanding) | Dataset characterisation, clinical-context mapping |
| | PP (Preprocessing) | Signal ingestion, filtering, resampling, artefact detection |
| | FEE (Feature Engineering) | Feature extraction across 10 extractor families |
| | MOD (Modelling) | Phase 1 (ML) and Phase 2 (DL) pipelines |
| | BAG (Brain Age Gap) | CN-BAG calculation, age-bias correction, outcome association |

| Tier | Module (ID) | Responsibility |
|---|---|---|
| | VAL (Validation) | Hypothesis testing, confounding checks, code quality |
| | EDA (Exploratory Analysis) | Data quality assessment, feature distributions |
| | VIS (Visualisation) | Performance plots, BAG distributions, SHAP maps |
| *III — Orchestration* | | |
| | EXP (Experimentation) | Pipeline orchestration, HPC job submission, experiment registry |

### C.1.3 Process Definitions (P1–P8)

Workflow execution followed eight formal process definitions to ensure reproducibility.

**Table C.2:** Standard operating procedures for major workflows.

| Proc. | Name | Scope |
|---|---|---|
| P1 | Requirements Eng. | Translate objectives into hypotheses and acceptance criteria. |
| P2 | Experiment Exec. | Assign EXP-NNN ID; log commit, split, and params; archive results. |
| P3 | Code Audit | Review module against register; verify test coverage. |
| P4 | Decision Making | Record design decisions with rationale in the Authoritative Document. |
| P5 | Documentation | Update existing documents in place; no parallel documents. |
| P6 | Issue Reporting | Assign unique ID; describe behaviour; link to module register. |
| P7 | HPC Job Submission | Validate on local test data before submitting to SLURM. |
| P8 | Session Start/End | Read control index/briefing; update registers and task statuses. |

## C.2 Engineering Infrastructure

### C.2.1 HPC Cluster Architecture

All experiments executed on the GWDG KISSKI allocation, which provides three SLURM partitions. The `kisski` and `kisski-h100` partitions provide GPU-accelerated compute nodes for Phase 2 deep learning, while the `jupyter` partition provides high-memory CPU nodes for interactive exploration and preprocessing.

**Table C.3:** GWDG KISSKI HPC partition summary. CPU architecture labels (Zen3, Zen4) are AMD EPYC generation designators. RAM and GPU figures are per-node.

| Partition | Nodes | CPUs | RAM | GPU | Notes |
|---|---|---|---|---|---|
| `kisski` | 24 | 128 (Zen3) | 512 GB | 1× A100 | Phase 2 training; GPU-accelerated; SLURM array jobs supported |
| `kisski-h100` | 8 | 192 (Zen4) | 1024 GB | 1× H100 | Phase 2 training; GPU-accelerated; `--gres=gpu:H100:1` |
| `jupyter` | 16 | 80 | 768 GB | — | Preprocessing, Phase 1 ML |

**GPU nodes (`kisski-h100`).** Phase 2 deep learning training used the `kisski-h100` partition with one NVIDIA H100 GPU (≈96 GB HBM3) per job. Mixed-precision training (PyTorch AMP) was enabled by default, reducing memory pressure and improving throughput on the H100 tensor cores. Each 5-fold training run occupied a single GPU for a low-double-digit number of GPU-hours, depending on the architecture, checkpointing behaviour, and whether only final training or also aborted comparison runs were counted.

**CPU preprocessing optimisation.** Preprocessing on `kisski` used `joblib.Parallel` with the `loky` backend to run the 10 Tier-2 feature extractors in parallel subprocesses, bypassing the Python GIL. With 16 CPUs per task (4 BLAS threads per loky worker), measured CPU utilisation remained at approximately 10% due to the sequential single-patient pipeline structure: ICA, filtering, and resampling execute serially per patient and dominate wall time. The `jupyter` partition achieved 87% utilisation via a per-segment GNU parallel approach (`EXP-001-gfn-preprocess.sh`), dispatching each (patient, segment) pair as an independent task across 40 concurrent workers on an 80-core node.

### C.2.2 GIL Bottleneck Resolution

See Section C.2.1 for the per-partition utilisation analysis.

### C.2.3 Storage and Data Management

Two parallel storage systems serve different partitions:

- **Lustre scratch** (`/scratch-grete/`): high-throughput parallel filesystem; accessible from `kisski` and `kisski-h100` partitions only. Used for all I-CARE raw data and preprocessing outputs.
- **VAST-NHR** (`/mnt/vast-nhr/`): universally accessible across all partitions including `jupyter` (GFN) nodes. Used as fallback and for jobs submitted to the `jupyter` partition, which cannot access `/scratch-grete/`.

The I-CARE corpus occupies 1.7 TiB raw. Preprocessing across Profiles A, B, C, and E expanded this to 5.8 TiB on the Lustre scratch filesystem. Feature matrices (Parquet) were stored with memory-mapped NPY signal files to eliminate RAM bottlenecks during Phase 2 training.

## C.3 Testing and Quality Assurance

### C.3.1 Quality Gates and No-Drop Policy

The project quality gates were designed as *diagnostic* controls rather than as cohort-pruning mechanisms. No patient was removed from the study solely because a gate exposed severe artefact load, a transformation failure, or an unexpected recording pathology. Instead, the affected epoch or segment was flagged, propagated as low-reliability, or excluded only from the specific downstream operation that required cleaner data, while the patient remained in the audited cohort.

This no-drop policy served two purposes. First, it prevented artificial performance inflation through selective cleanliness bias. Second, it made the gate outputs scientifically useful: the same telemetry that protected the pipeline during execution also documented where and why data quality deteriorated, allowing later sensitivity analyses and dashboard-based inspection without silently rewriting the cohort.

### C.3.2 Bugs Identified and Resolved

Five critical bugs were identified during development:

- **PyTorch thread over-subscription:** Fixed by pinning PyTorch to single-threaded execution (`torch.set_num_threads(1)`) during ICA fitting.
- **MNE ICA memory leak:** Resolved by adding explicit object deletion and `gc.collect()` on all early-exit paths.
- **ICLabel on unreferenced signal:** Introduced status checks to skip ICLabel when PREP robust referencing fails.
- **Burst-suppression hang:** Implemented a daemon thread wrapper with a timeout guard to prevent $O(n^2)$ search paths on raw cardiac artefacts.
- **NaN propagation:** Introduced a sanitisation step replacing non-finite source samples with per-channel medians before metric computation.

## C.4 Task and Pipeline Monitoring

### C.4.1 Task Governance Dashboard

The `task_dashboard.html` is a serverless single-page application that parses decentralised `00-TASKS.json` files from each module directory at load time, rendering a live Kanban board with status, priority, submodule tags, estimated hours, and dependency flags. No backend infrastructure is required. In this project, the dashboard was more than a convenience layer: it was a practical coordination instrument for keeping code, analysis, and thesis-writing work synchronised across a large modular workflow. The final state at thesis submission recorded 98 tasks to-do, 13 in-progress, 5 blocked, 105 done, and 27 won't-do across all project scopes.

### C.4.2 Module Data-Flow Dashboard

The `module_dashboard.html` is fed directly by the `EXPERIMENTS.md` single source of truth. It exposes a data-flow matrix showing which artefacts each pipeline module produces or consumes, the integration status of every EXP-ID, and control-plane operation summaries (Figures C.3 and C.4). Like the task dashboard, it functioned as

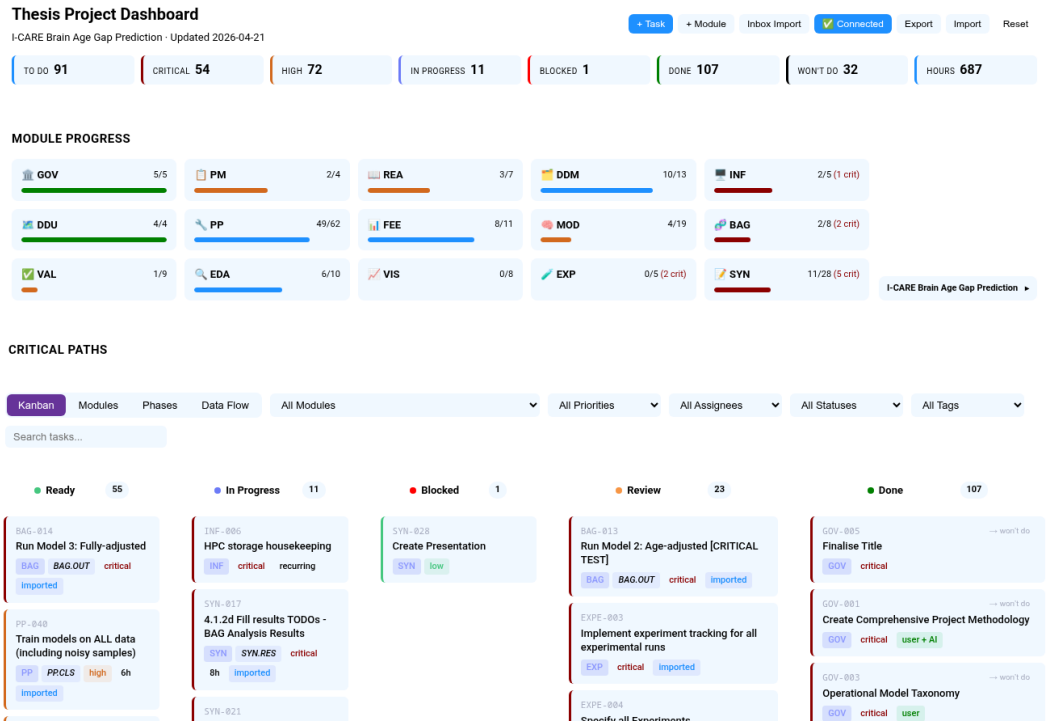
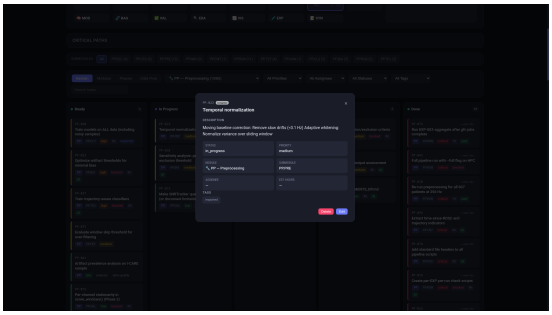
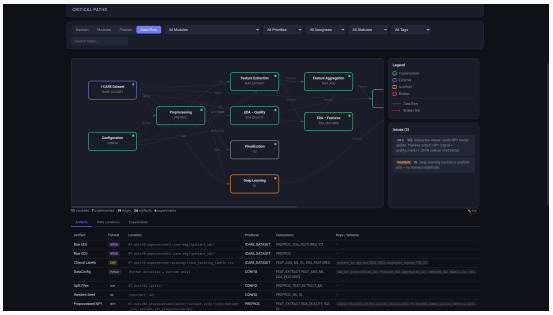



**Figure C.1:** The `task_dashboard.html` Kanban board showing task status across all modules. Cards carry status, priority, submodule tags, estimated hours, and dependency flags. The aggregate summary row reports totals at final thesis submission.



**(a)** Preprocessing module: expanded task card detail (In-Progress column)



**(b)** Alternative DAG view of module task distribution

**Figure C.2:** Two additional views of the task dashboard illustrating per-module filtering. The expanded card (left) shows the full metadata for the *Temporal normalisation* task; the right panel shows a different module context. Each tab is rendered client-side from the same `00-TASKS.json` source files, with no page reload required.

part of the project's reproducibility infrastructure: an inspectable architectural view that connected individual scripts and outputs back to the larger experimental system.

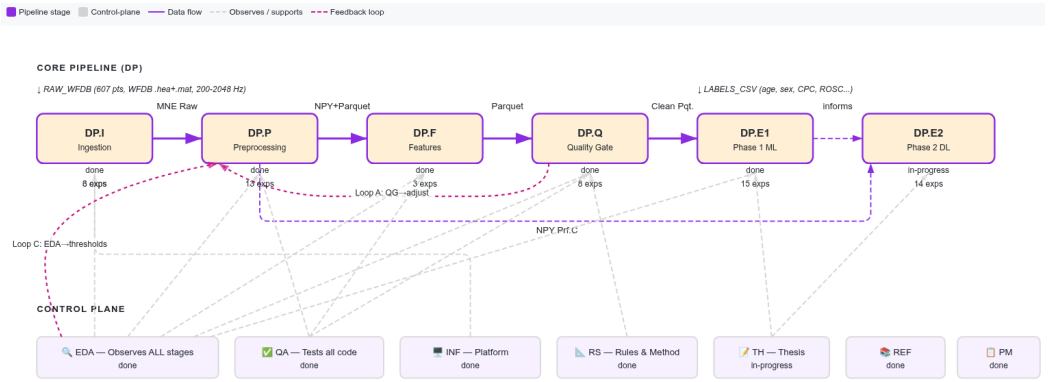



**Figure C.3:** Pipeline-flow tab of `module_dashboard.html` showing the six core pipeline stages (DP.I → DP.P → DP.F → DP.Q → DP.E1 → DP.E2) with data-flow edges, stage completion status, and feedback loops. The artefacts and data-locations tables enumerate the 26 pipeline artefacts with their producers, consumers, and directory paths.

### C.4.3 Pipeline Evaluation Dashboard

The `pipeline_evaluation_dashboard.py` (`04-code/05-visualization/`) is a Panel/Bokeh application for interactive quality control of the five preprocessing profiles across all 607 patients. It is invoked by pointing it at a pipeline output directory and exposes eleven tabs:

1. **Quality Summary** — per-patient, per-profile quality score table with sortable columns.

2. **Segments** — per-segment waveform and quality-mask detail for a selected patient and profile.

3. **SNR Tracking** — SNR trajectory across recording time and profiles.

4. **EECI** — EEG–ECG Contamination Index summary across profiles and patients.

5. **Pipeline Status** — pass/fail summary per pipeline stage.

6. **Feature Tier** — extracted feature statistics per tier (spectral, morphological, connectivity).

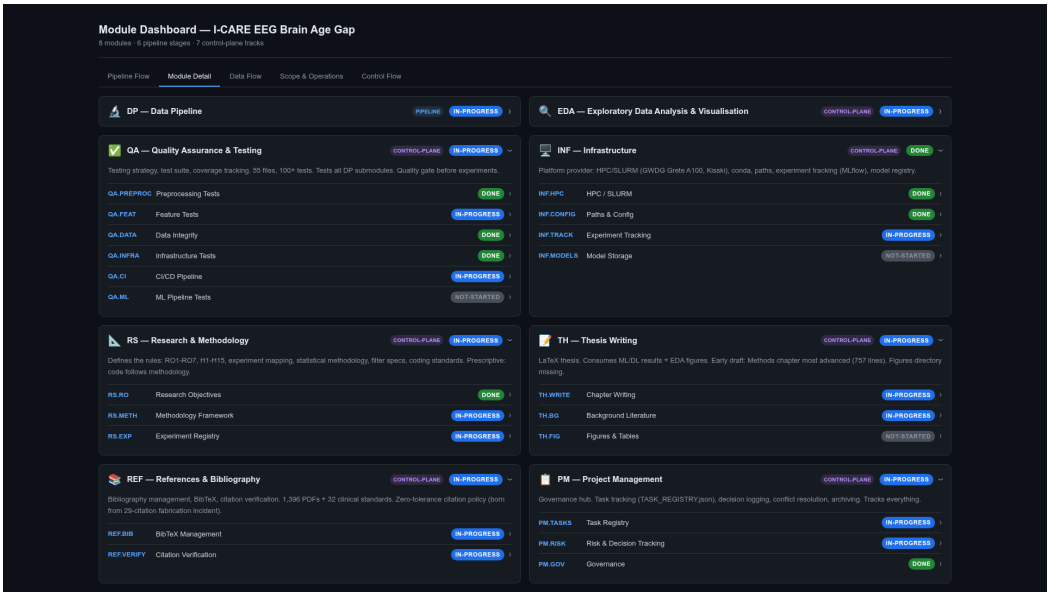



**Figure C.4:** Module-detail tab of `module_dashboard.html` showing per-module status cards for all 8 modules (DP, EDA, QA, INF, RS, TH, REF, PM) with submodule completion indicators. Submodule counts and status labels are auto-generated from the decentralised `00-TASKS.json` files.

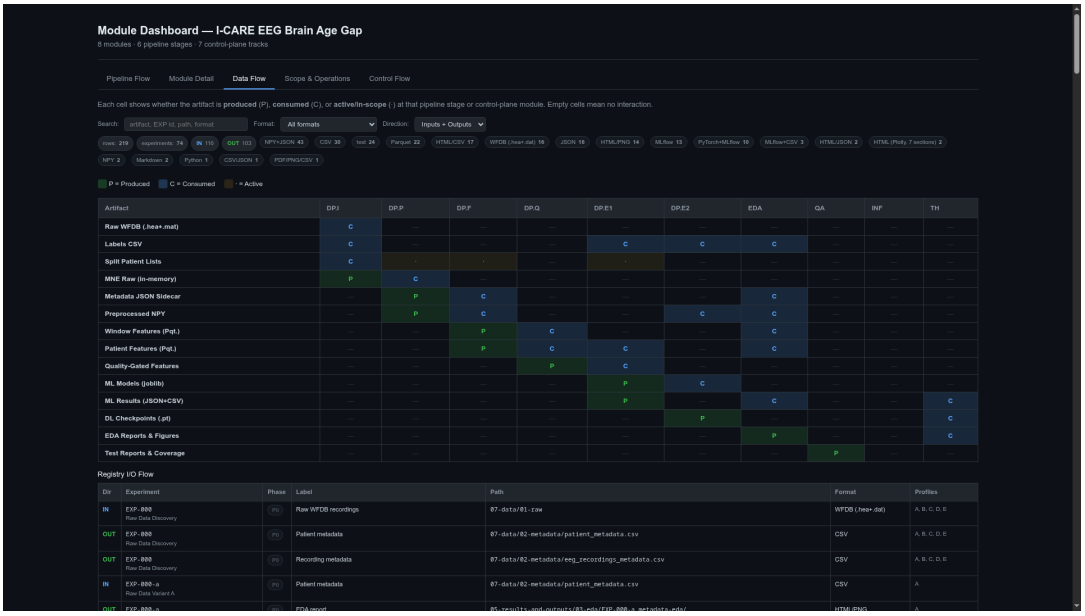



**Figure C.5:** The Module Dashboard data-flow view for the I-CARE EEG Brain Age Gap project. Rows represent data artefacts (raw WFDB, labels, preprocessed NPY, window and patient feature Parquets, ML/DL models, EDA reports); columns represent pipeline modules (DP.I through DP.E2, EDA, QA, INF, TH). Green cells indicate the module *produces* that artefact; blue cells indicate it *consumes* it. The view provides an immediate architectural overview of the complete end-to-end data lineage.

7. **Signal Viewer** — interactive raw-vs-processed waveform browser with channel selector and time zoom.

8. **PSD Comparison** — power spectral density overlay across profiles for a selected patient.

9. **Quality Mask** — temporal heatmap of the quality mask bitmask across all windows.

10. **ECG Correlation** — cross-channel ECG contamination correlation matrix.

11. **Detections** — timeline of artefact and pathology detections per window.

Figures C.6 and C.7 illustrate representative tabs rendered from patient-0284 data (variant A, raw/resampled profile).

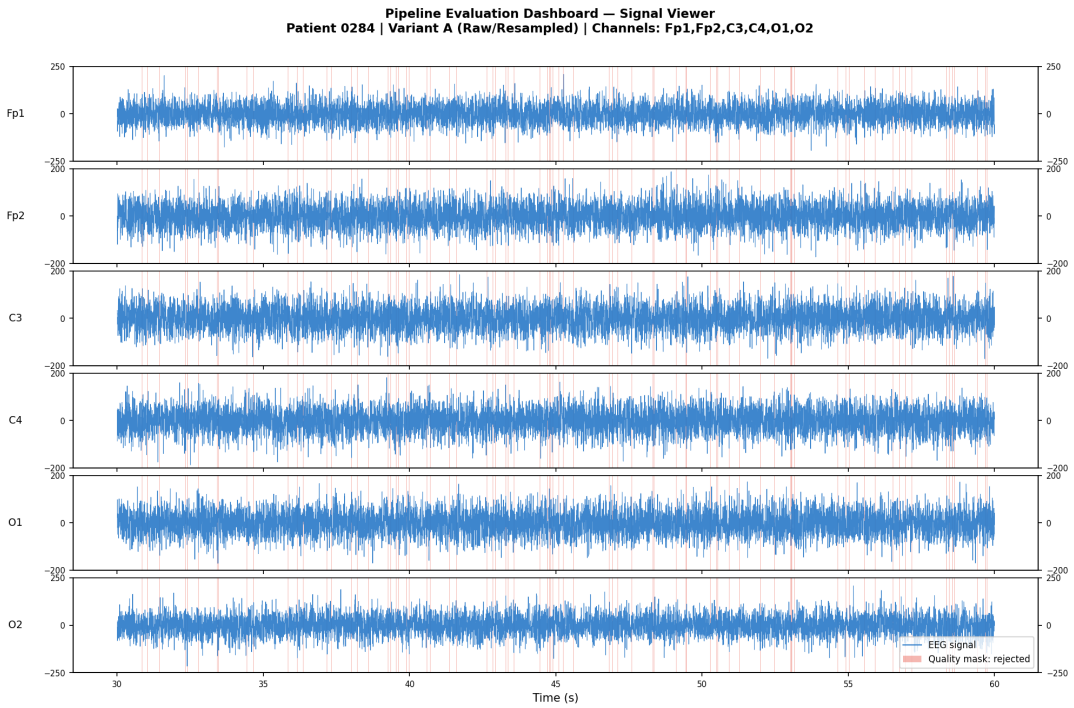
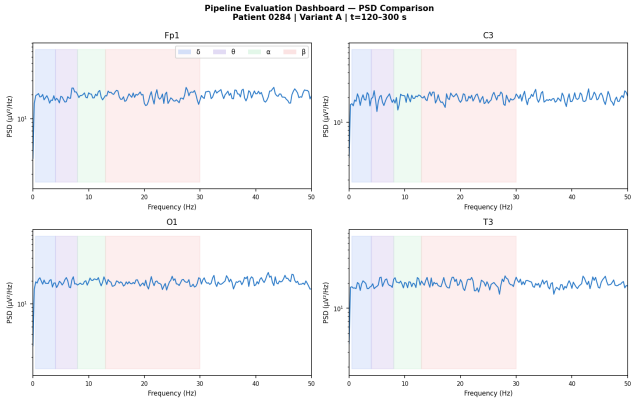
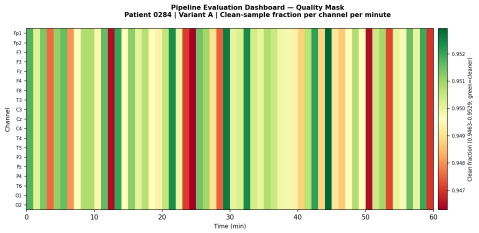



**Figure C.6: Signal Viewer tab.** Thirty seconds of six EEG channels (Fp1, Fp2, C3, C4, O1, O2) for patient 0284, variant A, starting at $t = 30$ s. Red hairlines mark individual samples flagged by the quality bitmask; unshaded samples are accepted. Flagged samples account for approximately 5 % of all samples in this segment and are short (1–3 samples each); the signal quality is uniformly high. Each channel is independently scaled to its peak amplitude.

These dashboards were built because no static report can convey the temporal structure of quality masks or the comparative impact of cardiac ICA removal across multiple




**(a)** PSD Comparison tab: Welch PSD for channels Fp1, C3, O1, T3. Coloured bands mark standard EEG frequency sub-bands ($\delta$, $\theta$, $\alpha$, $\beta$).




**(b)** Quality Mask tab: per-channel clean-fraction heatmap over the first 60 min of recording. Green = cleaner; red = more flagged (data range 0.946–0.953 across all channels and minutes, reflecting high overall signal quality).

**Figure C.7:** Two further tabs of the pipeline evaluation dashboard for patient 0284, Variant A. Both figures are rendered programmatically from the NPY output files using the same data pathway as the interactive dashboard.

profiles simultaneously. They therefore serve two roles: first, as essential working instruments during development and debugging; second, as substantive transparency artefacts for future researchers reproducing preprocessing decisions on the I-CARE dataset.

The same dashboards were also used to monitor job completion and stage-specific bottlenecks during production runs. What they showed most clearly was that execution cost depended strongly on what was being counted: full preprocessing and feature extraction were dominated by long CPU-bound stages, whereas fold-level Phase 1 fitting on already aggregated matrices was comparatively cheap, and Phase 2 costs depended on architecture choice plus reruns. For that reason, this appendix avoids a single thesis-wide runtime total and instead preserves the dashboard views as the more honest record of where time and compute were spent.

## C.5 AI-Assisted Research Workflow

This whole was not feasible without AI support.

Generative AI tools (large language models and retrieval-augmented generation systems) were integrated into the research workflow under continuous human oversight. The researcher's role was that of system architect and AI orchestrator: defining goals, evaluating outputs, identifying flaws, and retaining final authority over all scientific decisions. The following roles were defined:

**Conceptual discussion and refinement.**
LLMs served as interactive sparring partners to critically discuss, evaluate, and stress-test concepts, methodological outlines, and hypotheses that the researcher independently originated.

**Domain comprehension and literature review.**
RAG systems accelerated cross-disciplinary understanding and assisted with structured literature retrieval. The project codebase and documentation were ingested into NotebookLM notebooks via an adapted version of Pyragify [90], enabling grounded, hallucination-free queries about the project's own implementation and experiment plans through a local MCP server [91].

**Software engineering.**
LLMs assisted with the design and development of pipeline source code, debugging, testing, and refactoring. The researcher provided the conceptual design; all generated code was verified via automated tests (Section C.3).

**Project management.**
AI tools supported workflow organisation, including synthesis of research notes, task automation, and progress monitoring.

**Typesetting and proofreading.**
LLMs assisted with programmatic data extraction, formatting of LaTeX tables and figures, and extensive stylistic proofreading. No scientific text was generated entirely by AI without substantial human drafting and rewriting.

The final scientific authority remained exclusively with the human researcher. AI contributions were restricted to operational support, with all clinical claims and mathematical formulations subjected to manual primary-source verification. No section of this thesis was generated as a black-box output; instead, AI was used to accelerate the iteration of human-authored drafts and the extraction of insights from the local experimental corpus.

### C.5.1 Literature Provenance and Citation Gate

The literature workflow applied an MCP-first provenance rule for scientific claims. Before any literature-backed statement was inserted into notes or thesis text, the relevant NotebookLM MCP servers were queried against the local scientific corpus, and the

resulting answer was archived as a dated query record. This ensured that later synthesis could be traced back to a concrete retrieval event rather than to unlogged model memory.

The citation pipeline then enforced a strict human gate. AI tools were permitted to extract candidate paper metadata from MCP responses, map the result into project notes, and log newly surfaced sources as `TO VERIFY` in the verification tracker. They were not permitted to add entries to the authoritative bibliography. A paper entered `references.bib` only after independent human confirmation of authorship, title, year, and venue against the primary source or an already-verified local record.

This separation was introduced after an earlier internal citation-fabrication failure in November 2025 and remained mandatory throughout the thesis. In practical terms, the workflow had four non-negotiable steps: retrieve and archive the MCP response; synthesise the answer into notes; log any not-yet-verified paper in the tracker; and require human verification before bibliography entry. The purpose was not merely administrative neatness, but a reproducible audit trail for every scientific claim supported by AI-assisted retrieval.

### C.5.2 Project Toolchain Summary

**Table C.4:** Tools and systems used across the full project lifecycle.

| Tool / System | Category | Role |
|---|---|---|
| Python 3.11 | Language | All pipeline, modelling, and analysis code |
| MNE-Python | Signal proc. | EEG loading, filtering, ICA, montage |
| wfdb | I/O | WFDB-format I-CARE archive ingestion |
| XGBoost / scikit-learn | ML | Phase 1 gradient-boosted and linear models |
| PyTorch | Deep learning | Phase 2 CNN-BiLSTM-Attention architecture |
| MLflow | Experiment tracking | Run logging, metric archiving, reproducibility |
| SLURM | HPC scheduler | Job submission, array parallelism, resource tracking |
| Git + GWDG GitLab | Version control | Code versioning, CI/CD runner, commit-hash injection |
| NotebookLM + MCP | RAG / literature | Grounded literature queries; hallucination control |
| GitHub Copilot (Claude Sonnet) | LLM assistance | Conceptual discussion, code review, drafting |
| Pyragify | Corpus ingestion | Converting codebase and docs into NotebookLM source chunks [90] |
| LaTeX + Biber | Typesetting | Thesis composition and bibliography management |
| VS Code | IDE | Editing, Remote SSH, MCP server orchestration |

# Appendix D

# Formal Bayesian Proof of the Self-Fulfilling Prophecy

This appendix presents the formal Bayesian derivation underlying the self-fulfilling-prophecy argument that motivates the thesis's decision-independent framing. It defines the model variables, derives the observed-outcome function, decomposes the challenge metric, introduces the self-fulfilling-prophecy bias coefficient, and works through scenario analyses to the final methodological conclusion.

## D.1 Variable Definitions

The following variables define the interaction between the predictive model, the biological reality, and the clinical decision:

$M = 1$
: The machine learning model predicts a poor outcome.

$O = 1$
: An observed poor outcome (death or CPC 3–5) is recorded.

$W = 1$
: Withdrawal of Life-Sustaining Therapy (WLST) was performed by clinicians.

$N = 1$
: Irreversible brain injury (latent biological reality: irrecoverable).

$N = 0$
: Recovery potential exists (latent biological reality: recoverable).

## D.2 The Outcome Function

The observed clinical outcome $O$ is not a pure reflection of the biological state $N$, but a composite function of biology and clinical action:

$$O = N \vee W \tag{D.1}$$

This relationship implies that a poor outcome ($O = 1$) is observed if the patient is biologically irrecoverable ($N = 1$) OR if care is withdrawn ($W = 1$). Conversely, a good outcome ($O = 0$) can only be observed if the patient is both biologically recoverable ($N = 0$) AND care is maintained ($W = 0$).

## D.3 The Challenge Metric and Its Decomposition

Prognostic models are traditionally evaluated using sensitivity at a fixed false-positive rate. In the I-CARE context (PhysioNet Challenge 2023), the metric ($S$) targets sensitivity on observed poor outcomes:

$$S = P(M = 1 \mid O = 1) \quad \text{subject to} \quad P(M = 1 \mid O = 0) \leq 0.05 \tag{D.2}$$

By applying the law of total probability and the outcome function, we can decompose the observed sensitivity:

$$P(M = 1 \mid O = 1) = \frac{P(M = 1, N = 1) + P(M = 1, N = 0, W = 1)}{P(N = 1) + P(N = 0, W = 1)} \tag{D.3}$$

The second term in the numerator, $P(M = 1, N = 0, W = 1)$, represents the self-fulfilling prophecy. It counts instances where a model incorrectly predicts a poor outcome for a biologically recoverable patient ($N = 0$), but the prediction is arithmetically rewarded as a "true positive" because care was subsequently withdrawn ($W = 1$).

## D.4 The SFP Bias Coefficient

We define the SFP bias coefficient, $\gamma$, as the proportion of observed poor outcomes attributable to WLST in recoverable patients:

$$\gamma = \frac{P(N = 0, W = 1)}{P(O = 1)} \tag{D.4}$$

The relationship between the observed sensitivity ($S_{obs}$) and the true biological sensitivity ($S_{bio}$) is:

$$S_{obs} = (1 - \gamma) \cdot S_{bio} + \gamma \cdot P(M = 1 \mid N = 0, W = 1) \tag{D.5}$$

If a model employs features that clinicians also use to guide withdrawal (e.g., burst-suppression), the term $P(M = 1 \mid N = 0, W = 1) \to 1$. In such cases, a high reported sensitivity ($S_{obs}$) reflects a classifier of clinician behaviour rather than a predictor of biological recovery potential.

## D.5 Scenario Analysis

Table D.1 demonstrates how standard metrics treat each combination of state and action.

**Table D.1:** Scenario analysis of the interaction between biology and clinical action. Row 5 represents the critical confounder: a biologically recoverable patient whose care is withdrawn is rewarded as a true positive by the metric, hiding the biological error.

| Scenario | Bio. state ($N$) | Action ($W$) | Outcome ($O$) | Metric | Bio. reality |
|---|---|---|---|---|---|
| Irrecoverable, cont. | $N = 1$ | $W = 0$ | Poor ($O = 1$) | True Positive | True Positive |
| Irrecoverable, WLST | $N = 1$ | $W = 1$ | Poor ($O = 1$) | True Positive | True Positive |
| Recoverable, surv. | $N = 0$ | $W = 0$ | Good ($O = 0$) | True Negative | True Negative |
| Recoverable, mispred. | $N = 0$ | $W = 0$ | Good ($O = 0$) | False Positive | False Positive |
| **Recoverable, WLST** | **N = 0** | **W = 1** | **Poor ($O = 1$)** | **TP (Rewarded)** | **FP (Hidden)** |

## D.6 Methodological Conclusion

The presence of $\gamma > 0$ introduces an irreducible confound in any model trained to predict $O$ directly. This justifies the decision-independent strategy of this thesis: by training on

chronological age (a target that clinicians cannot change), this forces the model to learn from a signal that is mathematically decoupled from the clinical decision pathway.

# Appendix E

# Statistical Analysis Plan and Power Analysis

This appendix defines the statistical analysis plan registered prior to final analysis. Section E.1 documents the hypothesis register and testing framework. Sections E.1.1 and E.1.2 define the multiple-comparison strategy and within-hypothesis evaluation rules. Section E.2 provides power analysis and sample-size justification. Section E.4 presents the severity-score validation framework.

## E.1 Hypothesis Register and Testing Framework

Defining all hypotheses prior to final analysis guards against post-hoc reasoning inflating apparent significance in a dataset of this size. All hypotheses were formally defined, grouped by research objective, and tested at nominal $\alpha = 0.05$ with 95% confidence intervals throughout. The full RO-to-hypothesis crosswalk is given in Appendix Appendix B.

### E.1.1 Multiple Comparison Correction

To control the family-wise error rate across the nested hierarchy of experiments, we apply the following corrections:

- **Family F1 (H-1–H-4):** Bonferroni-Holm correction ($k = 4$) within the primary CN-BAG outcome family.
- **Sex Classification Control (H-5):** Standalone positive control (sex classification); no family correction applied; tested via permutation.

- **Family F2 (H-6, H-8, H-9):** Bonferroni-Holm correction ($k = 3$) within the preprocessing/representation family. H-7 (severity validation) is reported without family-wise correction as it is a diagnostic rather than a primary claim.
- **Exploratory (H-10, H-11):** Reported without family-wise correction; results are hypothesis-generating only.

### E.1.2 Evaluation Plan Within Hypotheses

Family-wise correction determines how claims are interpreted across hypotheses. Within each hypothesis, the test statistic is chosen to match the structure of the data rather than to maximise apparent sensitivity. Table E.1 summarises the guiding logic.

This logic follows the broader EEG outcome and decoding literature, where patient-grouped validation is used for segmented recordings, logistic outcome models are reported for post-anoxic prognostication, repeated-measures structures are retained for dynamic EEG analyses, and empirical nulls are preferred for decoder significance [70, 73–75, 86]. For paired preprocessing or modelling contrasts, the same logic applies: the unit of comparison is the patient or recording seen under both conditions, not the pooled windows [92, 93].

**Table E.1:** Test-selection rationale within the evaluation plan.

| **Analysis type** | **Statistic** | **Why this test is used** |
|---|---|---|
| Paired profile or model comparisons | Wilcoxon signed-rank | The same patients contribute observations under both conditions, so the pairing must be preserved. Error distributions are not assumed Gaussian, making a rank-based paired test preferable to a paired $t$-test. |
| Outcome association analyses | Spearman correlation; logistic regression | The main question is whether CN-BAG tracks outcome monotonically and whether it shifts the odds of poor outcome at patient level. Rank-based association is robust to non-normality, while logistic regression matches the binary CPC framing and yields interpretable odds ratios. |
| Grouped or repeated-measures analyses | Mixed-effects models | Windows and recordings are nested within patients and sites, so independence cannot be assumed. Mixed-effects models retain that structure instead of treating repeated observations as separate subjects. |
| Feasibility and diagnostic checks | Permutation tests; null MAE thresholds | Grouped cross-validation, class imbalance, and custom error metrics create null distributions that are not well captured by simple asymptotic formulas. Empirical nulls therefore provide the more defensible reference. |
| Independence from severity markers | Partial Spearman, Jaccard@20, incremental $R^2$ | These analyses ask whether CN-BAG remains distinct from established severity markers, not just whether it predicts well. They therefore target conditional association, feature overlap, and incremental explanatory value. |

## E.2 Power Analysis

Statistical power was estimated for the hypotheses with explicit power calculations using the 483-patient development set (Table E.2).

**Table E.2:** Statistical power for selected hypotheses ($N = 483$ development set, $\alpha = 0.05$).

| H | Test | Effect size | Power |
|---|---|---|---|
| H-6 | Wilcoxon signed-rank | $d = 0.2$ | 97.0% |
| H-1 | Correlation test | $r = 0.5$ | >99% |
| H-2 | Logistic regression (Wald) | OR = 1.5 | 98.0% |
| H-3 | Spearman correlation | $\rho = 0.3$ | >99% |
| H-4 | Mixed-effects ($N \approx 303$) | $r = 0.2$ | 86.8% |
| H-5 | Permutation test | — | >99% |
| H-10 | Wilcoxon signed-rank ($N \approx 150$) | $d = 0.3$ | 57.6% |

The sex-stratified hypothesis is designated exploratory due to lower power (57.6%). The remaining reported calculations exceed 85% power; hypotheses without entries in Table E.2 are handled descriptively or in the appendix B crosswalk.

## E.3 Null-Model Feasibility Thresholds

The MAE criterion for H-1 is formulated relative to the *constant predictor* (mean-age predictor), which assigns $\hat{y}_i = \bar{y}$ to every patient. Its expected MAE is the mean absolute deviation of the cohort age distribution:

$$\mathrm{MAE}_{\text{null}} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}| \tag{E.1}$$

For a normally distributed age variable with standard deviation $\sigma$, the asymptotic expectation is $\mathrm{MAE}_{\text{null}} \approx \sigma\sqrt{2/\pi}$. Because the null-model MAE depends on the age distribution of the analysis cohort, the threshold differs between the full development set and any subgroup:

- **Phase 1 training cohort** ($N = 386$, $\bar{y} = 60.8$ y, $\sigma = 15.6$ y): $\mathrm{MAE}_{\text{null}} = 12.61$ y. (This is the Phase 1 training subset of the 483-patient development set; the 97-patient temporal holdout is excluded here.)
- **CPC 1–4 subgroup** ($N = 162$, $\bar{y} = 58.1$ y, $\sigma = 13.7$ y): $\mathrm{MAE}_{\text{null}} = 10.92$ y.

The one-sided permutation threshold is the $5^{\text{th}}$ percentile of the MAE distribution obtained by shuffling age labels 1,000 times and computing the patient-level MAE of the mean-age predictor in each permutation. For the Phase 1 training cohort, this threshold is $12.54$ y. The pre-specified success criterion of MAE $< 12$ y is set conservatively below this permutation threshold to provide a margin above sampling variability in the null estimate; a model achieving exactly $12.5$ y would not constitute convincing trait recovery.

## E.4 Severity-Score Validation

Before interpreting any CN-BAG outcome association, the severity-score validation protocol tests whether CN-BAG is merely a proxy for established EEG severity markers rather than an independent biomarker. Three pre-specified statistics are computed:

1. **Partial Spearman $\rho$:** CN-BAG $\leftrightarrow$ CPC, controlling for burst-suppression ratio, inter-burst interval, continuity index, and ROSC time.
2. **Feature overlap (Jaccard@20):** The top-20 features of the age model are compared with the top-20 features of a CPC severity model. Jaccard $\geq 0.7$ indicates CN-BAG is a severity proxy; Jaccard $< 0.7$ indicates a distinct feature profile.
3. **Incremental $R^2$:** $R^2$ of $\texttt{outcome} \sim \texttt{CN-BAG}$ versus $\texttt{outcome} \sim \texttt{EEG severity features alone}$.

# Declarations

## Ethical Disclaimer

This thesis is a retrospective secondary analysis of de-identified data from the publicly available I-CARE/PhysioNet resource. No new patient recruitment, intervention, or clinical decision-making was performed as part of this work.

### Declaration on the Use of AI Assistance

Generative AI tools (large language models, retrieval-augmented generation systems) were integrated into the research workflow under continuous human oversight. The core conceptual work, intellectual direction, and scientific conclusions remain exclusively my own. I retained scientific control, ethical responsibility, and final authority over all methodological decisions, data interpretations, and conclusions.

In this work I have used AI (1) during brainstorming; (2) when creating the outline; (3) to write individual passages, altogether to the extent of approximately 40% of the entire text[1]; (4) for the development of software source code; (5) for optimising or restructuring software source code; (6) for proofreading or optimising. Furthermore (A) for literature retrieval and synthesis via RAG systems; (B) for statistical reasoning support; (C) for batch analysis of clinical data; (D) for the creation and enhancement of figures, tables and design; (E) for bibliography management and gap analysis.

The methodological integration of these tools is described in Section C.5. Detailed tool-to-task mappings, the complete governance rules, and explicit exclusions are documented in Sections C.1 and C.5.

[1] This figure estimates the proportion of text for which AI produced an initial draft subsequently revised by the author. All intellectual content, methodological decisions, and scientific conclusions are exclusively the author's.

## Statement of Originality

I hereby declare that this thesis is my own work and that all sources and aids used have been acknowledged.