International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-14, Issue-6, Nov-Dec 2026
DOI: https://doi.org/10.55524/ijircst.2026.14.6.8
Article ID IRPC-1437, Pages 64-90
www.ijircst.org

Bridging Heart and Brain through Multimodal ECG-EEG Deep Learning

Michidmaa Arikhad[1], *Saad Rasool[2], and Gnanesh Methari[3]

[1, 2] Department of Computer Science and Technology, American National University, Louisville Kentucky, United States
[3] Department of Information Technology (Cybersecurity), Franklin University, Columbus, United States

*Correspondence should be addressed to Saad Rasool; saadrasool6@gmail.com

Received: 16 Oct 2026; **1st Revision:** 30 Oct 2026; **2nd Revision:** 10 Nov 2026;
Accepted: 22 Nov 2026; **Published Online:** 29 Nov 2026

Copyright © 2026 *Saad Rasool et al. This is an open-access article distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT- There is a coupling between the heart and the brain via the autonomic nervous system and many clinical and consumer recordings involve both an electrocardiogram (ECG) and an electroencephalogram (EEG). Multimodal deep learning combining the two modalities is appealing, but current studies of multimodal systems almost never check the added value of the second signal over the best single signal, do not validate across multiple subjects and datasets, and almost never examine what happens when one of the signals is missing or corrupted. In this paper we specify a pre-registered experimental programme which does all three. We compare seven models, ranging from a classical feature extraction baseline through single-modality networks to early, late and cross-attention fusion, and a modality dropout contrastively pre-trained bridge model, on three problems: sleep staging, arousal detection and affect recognition. The validation is done subject-disjointly, involves a large polysomnography dataset with two external datasets and is framed in terms of four metrics: the gain in complementarity over the best single-modality model, the retention of performance when one signal is missing, the loss in performance when going from one dataset to another, and calibration and subgroup fairness. Modality reliance is measured by an integrated gradients-based modality reliance index. Sample size calculation is done for a paired test over per-subject differences, minimum detectable effects are specified for each problem, all thresholds, budget of hyperparameter tuning and decision rules are fixed in advance of opening the test set. The programme is a research design and not the evaluation of any diagnostic device. Results are not reported in the present version, tables are prepared but not populated yet.

KEYWORDS- Multimodal Deep Learning; Electrocardiogram; Electroencephalogram; Brain-Heart Interaction; Sleep Staging; Contrastive Learning; Missing Modality; Calibration; Fairness; Healthcare Artificial Intelligence

I. INTRODUCTION

Neither the heart nor the brain can be considered in isolation from each other. Central autonomic networks determine cardiac rhythm, while cardiac rhythm, in its turn, reflects the state of the brain: the heart-rate variability is different during various phases of sleep, transients in arousal are followed by heart rate increase, and problems with the one organ lead to problems with the other one – just like cardiac disorders occsssssssssurring after a stroke [1], [2], [3]. As the two measurements are related, combining them into a single record can presumably provide more information about the state of a person than measuring just any of the signals.

However, this potential has not been explored to its full extent. With deep learning methods, it became possible to classify cardiac rhythms based on ECG signals at a level of cardiologists' expertise [4], [5]; EEG was effectively decoded in the end-to-end manner [6], [7], and for the case of sleep, there appeared machine learning models matching the performance of human scorers on EEG data [8], [9], [10]. There also exist studies combining the two signals, but mainly in affective computing, due to availability of datasets featuring both signals [11], [12]. Nevertheless, such works are rather few, and are validated on a single dataset, with comparisons made against weak single-signal models.

Practical concerns make the issue even more interesting. First, a fused model should outperform the best single-signal model in order to be worth building, because a gain lower than the variance of performance due to random initialization is not significant. Second, in practice, one of the inputs will very likely be unavailable or will be noisy in some way, because an EEG electrode may become detached, an ECG lead will be noisy, or a device records only one signal. A fused model that stops working without one of the inputs will not be useful for real-life applications [13], [14]. Third, the performance measured on the dataset containing records of the same subjects in both training and testing set is overestimated; it is known as a common problem for clinical machine learning applications [15], [16].

Thus, we frame the multimodal ECG-EEG deep learning gain as a measurement question and subject it to a single pre-registered programme based on five questions. Is there an improvement over the best single-signal model beyond seed noise (RQ1)? How does the answer depend on the task (RQ2)? Does the improvement resist missing/ degraded signal and cross-dataset testing (RQ3)? Are the

fused models calibrated, subgroup-equitable, and properly utilising both inputs (RQ4)? And is the gain worth the cost in comparison with the strong classical baseline (RQ5)?

There are four key contributions of this paper. First, it defines the measurements answering these questions: complementarity gain over the best single-signal model, missing-modality retention, the cross-dataset drop, and modality reliance index. Second, it pre-specifies the set of models, the tuning budget, splits, stress tests, decision criteria, and gives equal tuning opportunities to every model so that the gain cannot come from the additional tuning of a fused model. Third, it introduces a contrastively pre-trained model with modality dropout as another candidate method and compares it to simpler fusion approaches and a classical baseline to measure the added complexity value. Fourth, it pre-specifies how each possible result should be interpreted, including those where the second signal brings no added value.

The programme is a research design. It does not evaluate a medical device, and none of its results can be used by itself to make a clinical claim. The rest of the paper is organised as follows. The related literature is reviewed in section 2. In section 3, we introduce the framework. The methodology is specified in section 4. In section 5, the pre-registered predictions and reporting structure are stated. In section 6, we fix the interpretation of each possible result. Section 7 is a conclusion. Figure 1 illustrates the concept and the working principle of the proposed program.

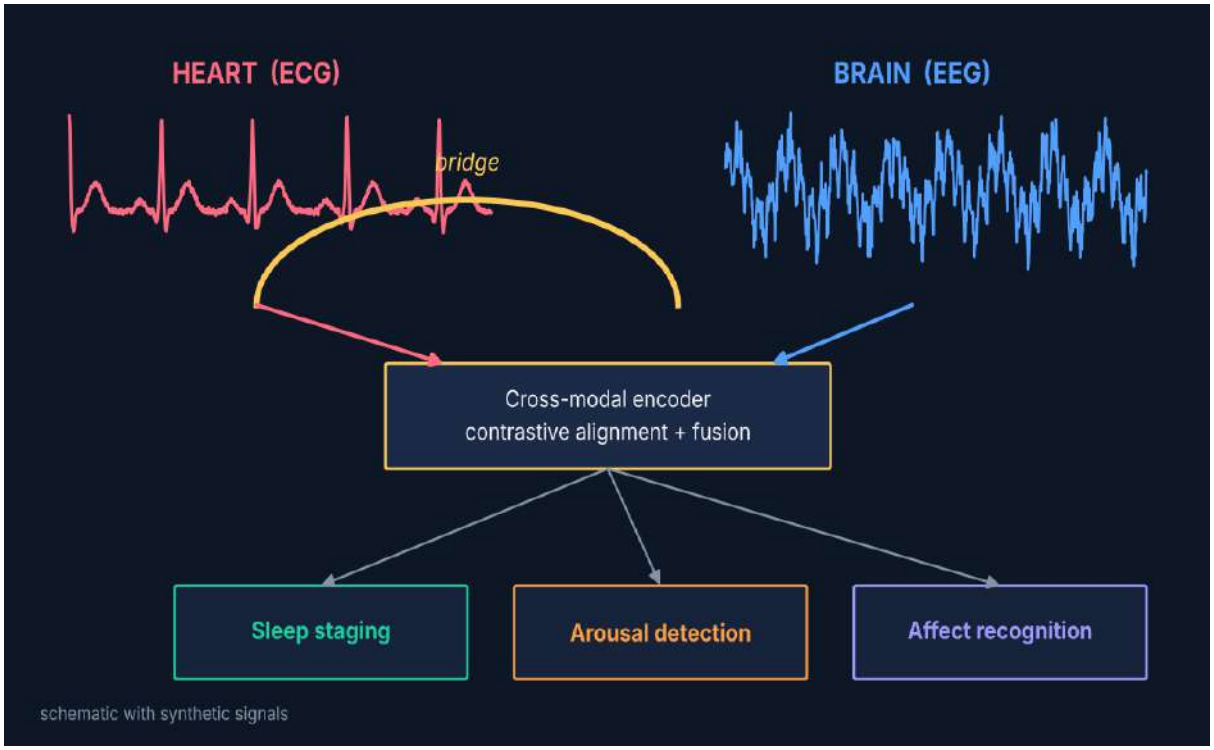

Figure 1: Heart and brain signals are encoded separately, aligned and fused, and the resulting representation is evaluated on three tasks. The signals shown are synthetic and for illustration only

II. LITERATURE REVIEW AND RELATED WORK

A. Brain–Heart Interaction and Heart Rate Variability

The physiological rationale for fusing the two signals lies in the autonomic nervous system. Review articles have described the regulation of cardiac autonomic output via the cerebral cortex and the brain stem as well as the feedback of the cardiac input to the brain [1], [2]. The clinical literature reports the effects of a break in the connection, including arrhythmia and myocardial damage secondary to stroke [3]. Heart rate variability (HRV), computed from the intervals between successive heart beats, is the conventional measure of autonomic tone, with time domain and frequency domain HRV analysis being delineated in a task force document [17](See the figure 2 and Table 1).

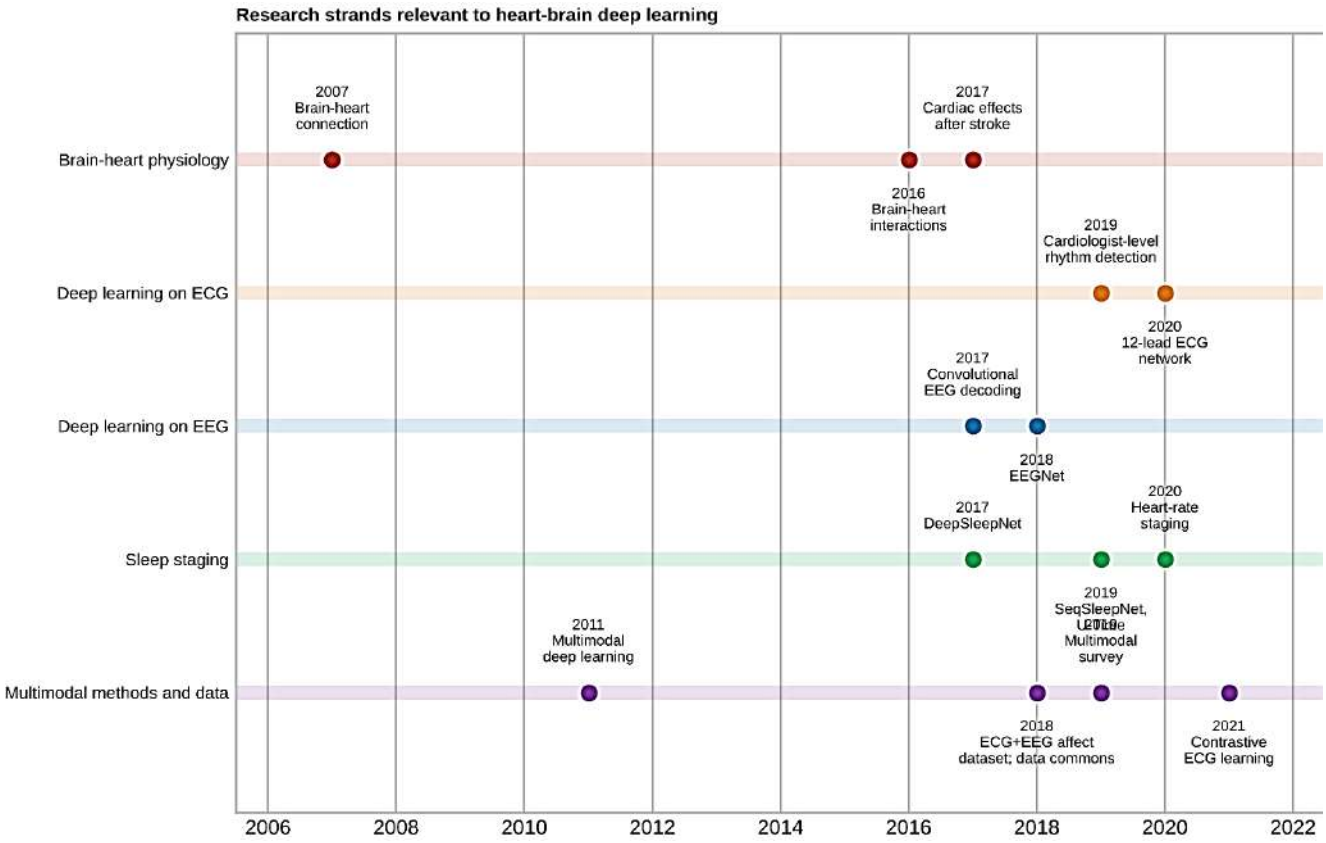

Figure 2: Research strands relevant to heart–brain deep learning, 2007 to 2021. Markers show the year of the cited work; the assignment of each work to a strand is a simplification.

Table 1: Strands of prior work and what each leaves open

Strand	What it established	Data and signals	What it leaves open
Brain–heart physiology	Central autonomic control links brain state and cardiac rhythm; cardiac complications follow brain injury [1], [2], [3]	Physiological and clinical studies	No benchmark of what a model gains by using both signals
Deep learning on ECG	Cardiologist-level rhythm classification [4], [5]	Large ECG datasets	Heart signal only
Deep learning on EEG	Compact networks decode EEG end to end [6], [7], [18]	EEG datasets	Brain signal only
Sleep staging and arousal	Staging from EEG, and from heart rate alone [8], [9], [19], [10]	Polysomnography cohorts	Signals studied separately; cross-dataset results vary
Multimodal methods and datasets	Fusion taxonomy; missing-modality methods; ECG+EEG affect datasets [20], [13], [11], [12]	Small multimodal datasets	Small samples; subject-dependent protocols common
Evaluation critiques	Record-wise splits and leakage inflate performance [15], [16]	Methodological analyses	Motivates subject-wise design here

B. Deep Learning on ECG and EEG

Convolutional nets used on unprocessed ECG achieve the same results as cardiologists in rhythm classification on big datasets [4], [5]. On EEG, end-to-end convolutional nets decode brain activity directly and illustrate what they rely on [6], and small networks with limited number of parameters work well across several paradigms [7]. A review of deep learning on EEG highlights the diversity of neural net architectures and, in particular, the variability of evaluation methods used in different studies [18]. Residual nets [21] are the standard architecture for one-dimensional physiological signals, and gradient-boosted decision trees on feature engineered data remain a solid baseline on tabular summaries [22].

C. Sleep Staging and Arousal Detection

Sleep is scored using 30-second epochs divided into wakefulness, three stages of non-rapid eye movement and rapid eye movement sleep according to established criteria [23], and brief arousals are scored using separate criteria [24]. Deep nets score sleep based on single channel EEG by modeling epochs and performing sequence prediction [8], [9], and by using fully convolutional segmentation

[19]. Heart rate alone is enough to stage sleep at a lower accuracy [10]. Two large publicly available datasets contain overnight polysomnography with both signals: a community cohort accessible via a sleep data common [25], [26], and the labeled dataset of a polysomnography challenge from PhysioNet [27], [28].

D. Multimodal Fusion, Missing Modalities and Contrastive Pretraining

Multimodal learning depends on where the modalities are integrated – early, late or in intermediate representations – and if they are aligned or complementary [29], [20]. Cross attention enables one modality to query another [30]. Models trained on data without any missing modalities fail when one of the modalities is missing at test time. Dropping modalities randomly during training makes models more robust [13], and methods specifically designed for extremely missing modalities have been proposed [14]. Contrastive objectives train aligned representation of signals by pulling together co-occurrences and separating dissimilar ones [31], and they have been successfully applied to cardiac signals in time, space and across patients [32]. These methods inspire the bridge model, yet their benefits have not been tested against simpler approaches on ECG-EEG tasks in a unified protocol.

E. Evaluation Pitfalls: Leakage, Calibration and Fairness

Fissioning windows from the same subject into training and test datasets improves the performance, as it gives the model subject-specific features which are needed when using such models in clinical applications [15]. Data leakage is among the main sources of non-reproducibility of machine-learning science in general [16]. Modern deep learning methods are known to be overconfident, and proper calibration must be provided [33]. Different subpopulations may have different levels of performance due to differences in sex, age, and ethnicity [34]; and guidelines for clinical predictions ask for performance on different subgroups and clear indication of use cases [35]. See the figure 3.

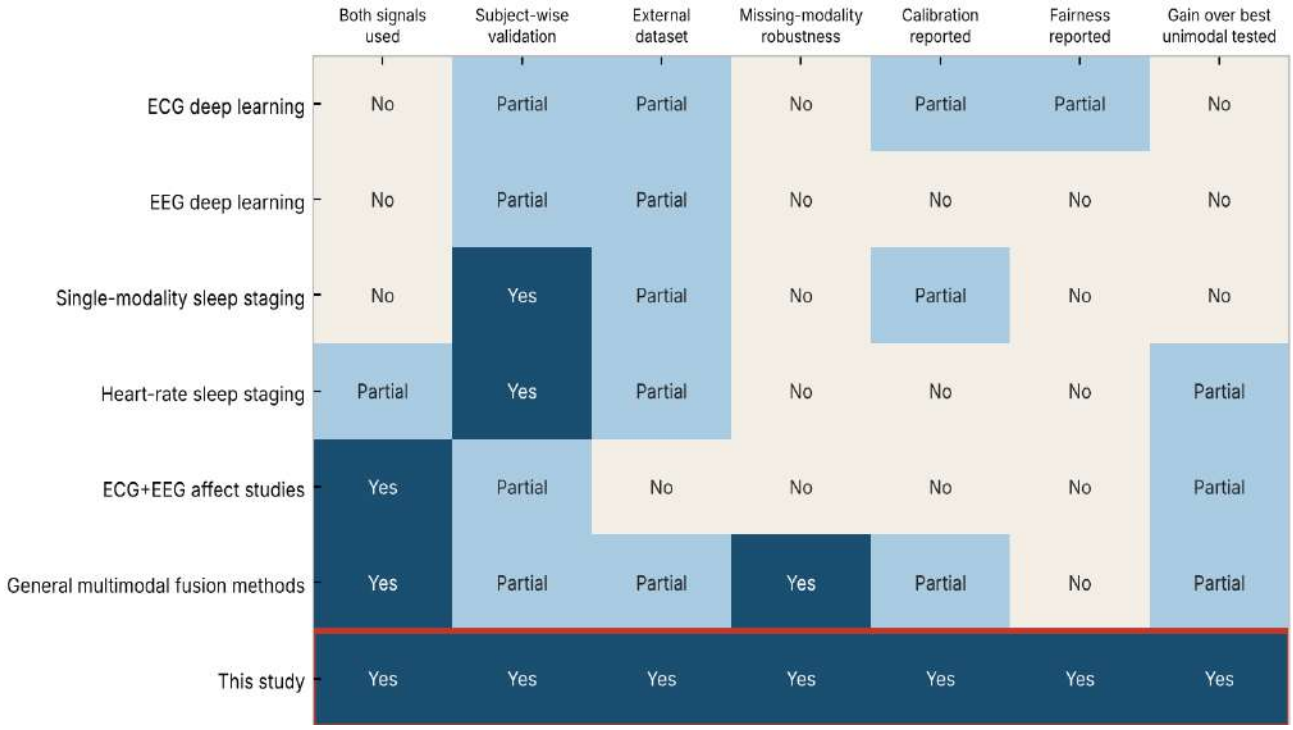

Figure 3: Coverage of seven requirements by six strands of related work and by the present study (No, Partial, Yes). The coding is a pre-registered starting hypothesis, re-coded from a literature scan.

Table 2: Positioning of the present study. The final row is the only entry that covers all seven requirements.

Strand of work	Both signals used	Subject-wise validation	External dataset	Missing-modality robustness	Calibration reported	Fairness reported	Gain over best unimodal tested
ECG deep learning	No	Partial	Partial	No	Partial	Partial	No
EEG deep learning	No	Partial	Partial	No	No	No	No
Single-modality sleep staging	No	Yes	Partial	No	Partial	No	No
Heart-rate sleep staging	Partial	Yes	Partial	No	No	No	Partial

ECG+EEG affect studies	Yes	Partial	No	No	No	No	Partial
General multimodal fusion methods	Yes	Partial	Partial	Yes	Partial	No	Partial
This study	Yes	Yes	Yes	Yes	Yes	Yes	Yes

III. THEORETICAL FRAMEWORK

A. Tasks and Datasets

Three tasks are selected since both the signals might provide information regarding them. In sleep staging, the brain signal is the traditional source of labels, whereas the heart rate is different during non-rapid-eye-movement and rapid-eye-movement sleep stages, which suggests that the heart rate might contribute information only where EEG signals are vague. In arousal detection, the arousal is a cortical event with an associated autonomic reaction, so these signals are expected to be complementary. In affect recognition, both brain and heart activities change according to the emotional state of a person; however, the datasets are limited, and the labels are subjective, which makes it an exploratory task. Table 3 provides the three tasks and their datasets, and Figure 4 shows the expected contribution of each signal for each task.

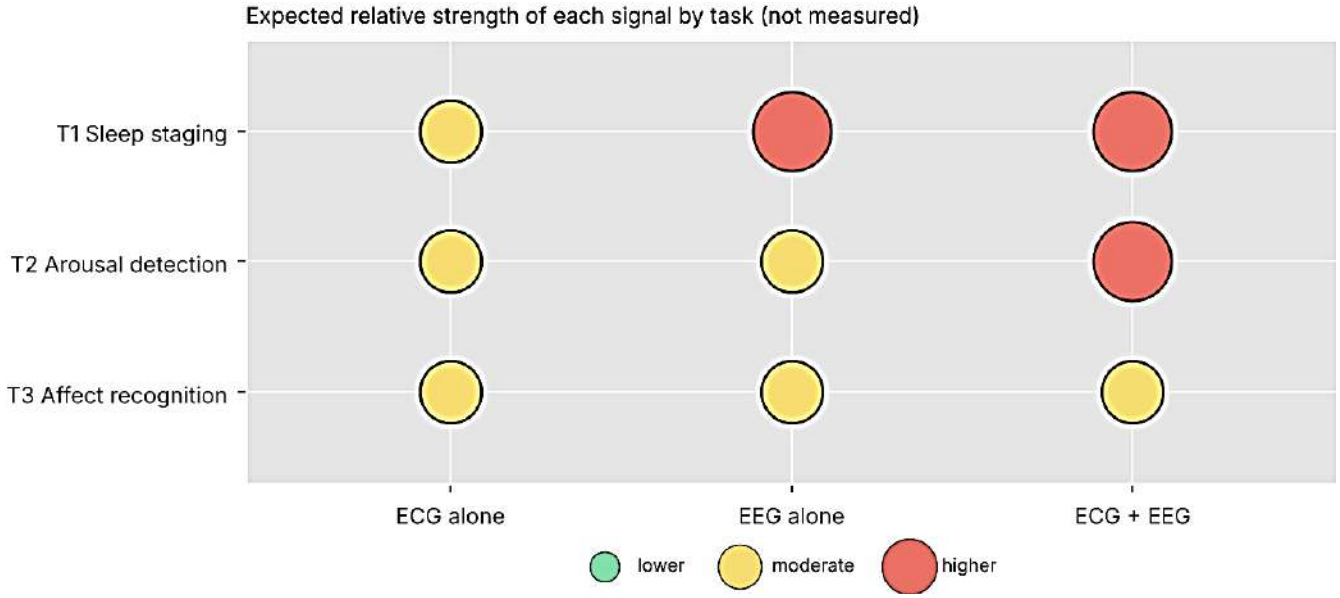

Figure 4: Expected relative strength of each signal alone and combined, by task. Values are expectations recorded before data collection and are not measured.

Table 3: Tasks and datasets. Counts and signal properties are planning values and are confirmed against the data dictionaries at access

Task	Label	Dataset	Signals	Role	Subjects
T1 Sleep staging	Five classes (wake, N1, N2, N3, REM) per 30 s epoch, scored under standard rules [23]	D1 community polysomnography cohort [25], [26]; D2 challenge polysomnography set [27], [28]	ECG and central EEG, resampled to a common rate	D1 train, validation and test; D2 external test	D1 about 5,800; D2 about 1,000
T2 Arousal detection	Binary per epoch: contains at least one scored arousal [24]	D1 and D2	ECG and central EEG	As T1	As T1
T3 Affect recognition	Binary valence and arousal from self-ratings, median split	D3 [11]; D4 [12]	ECG and 14-channel EEG	Subject-independent cross-validation; D3-D4 transfer	D3 23; D4 40

B. Models

Seven approaches are considered (Table 4, Figure 5). M0 is the classical approach: handcrafted features of heart rate variability and EEG band power features as the input for gradient boosted decision trees [17], [22]. Approaches M1 and M2 are deep learning unimodal systems and represent the best-unimodal baseline. Approaches M3, M4, and M5 perform early, late and cross-attention signal fusion respectively [30]. Approach M6 is the bridging approach;

it follows the same architecture as M5 but its two encoders are initialized through contrastive pretraining on co-recorded ECG and EEG windows [31], [32] and trained using the modality dropout [13]. All considered deep learning models have about 20% of the same parameter count of two million and equal context, hence the difference can be attributed to the utilization of the signals only.

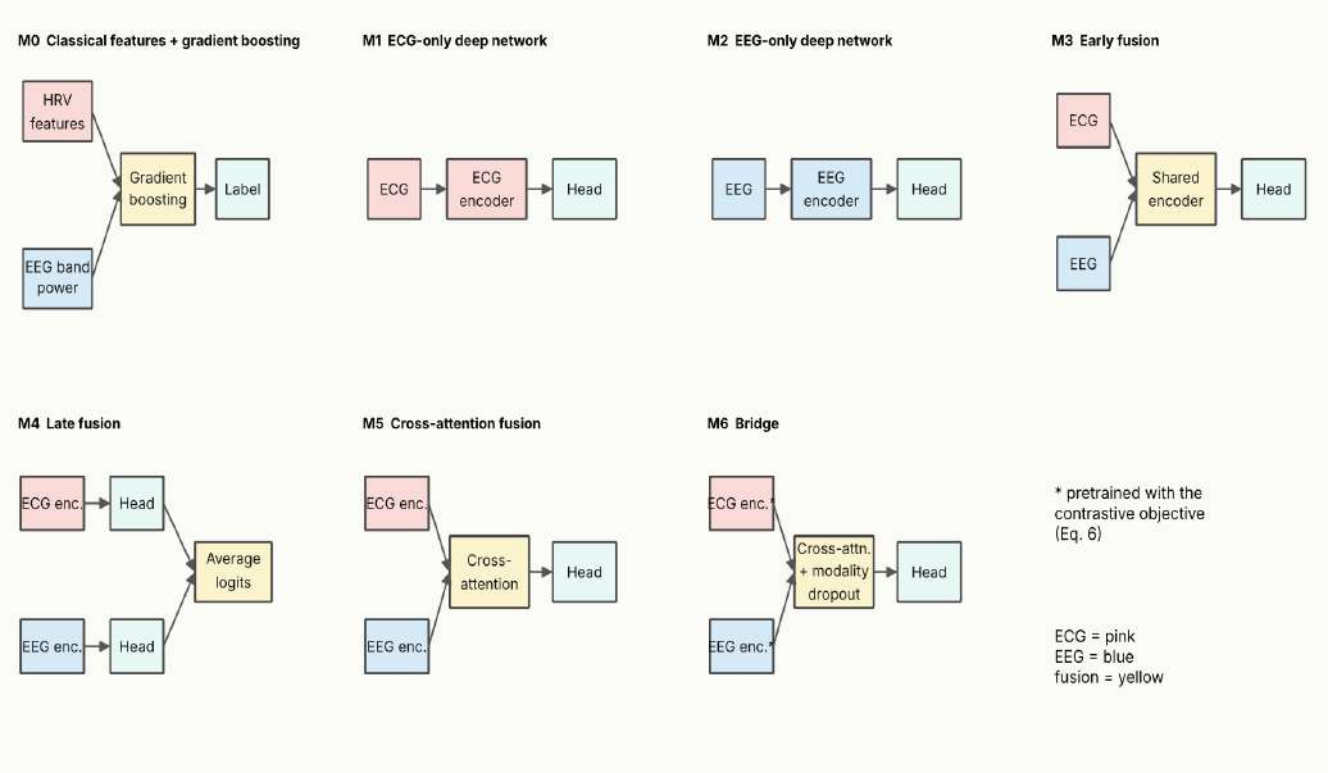

Figure 5: Schematic of the seven models. Asterisks mark encoders initialised by contrastive pretraining

Table 4: Models compared. Every model receives the same tuning budget

ID	Model	Inputs	How the signals are combined	Training
M0	Classical features + gradient boosting	HRV features; EEG band power	Concatenated feature vector	Boosted trees; 30-trial search
M1	ECG-only deep network	ECG	None	One-dimensional residual network [21]
M2	EEG-only deep network	EEG	None	Compact convolutional network [7]
M3	Early fusion	ECG and EEG	Channels stacked before a shared encoder	Supervised
M4	Late fusion	ECG and EEG	Separate networks; logits averaged	Supervised
M5	Cross-attention fusion	ECG and EEG	Token-level cross-attention between encoders	Supervised
M6	Bridge: contrastive pretraining + modality dropout	ECG and EEG	As M5, encoders pretrained to align the signals	Contrastive pretraining, then supervised with modality dropout

C. Research Questions

Mapping of research questions to answering phases, results, and analysis is shown in Table 5. Flow of evidence from research questions to phases and then results is illustrated in Figure 6.

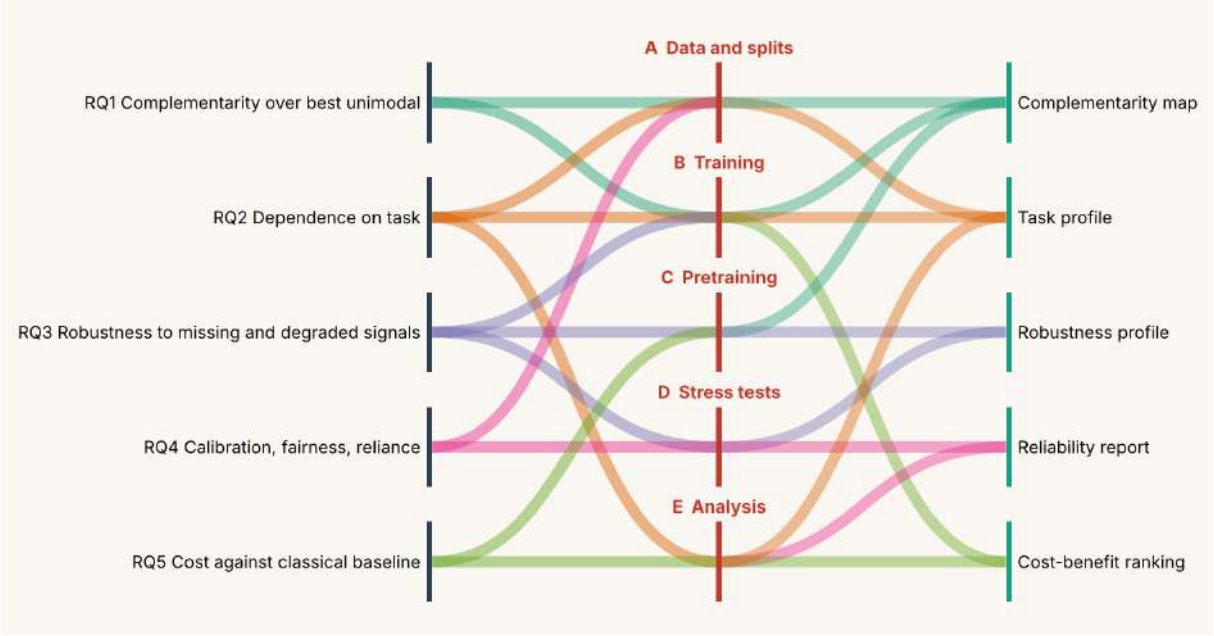

Figure 6: Flow of evidence from research questions through the five phases to the programme outputs

Table 5: Research questions, phases, primary outcomes and analyses

ID	Question	Phases	Primary outcome	Analysis
RQ1	Does a fused model improve on the best single-signal model by more than seed noise?	A, B, E	Complementarity gain CG	Paired test on per-subject differences
RQ2	How does complementarity depend on the task?	A, B	CG for T1, T2, T3	Task-wise contrasts; exploratory for T3
RQ3	Does the gain survive a missing or degraded signal and a new dataset?	C, D	Missing-modality retention MMR; cross-dataset drop CDD	Subject bootstrap intervals
RQ4	Are fused models calibrated, fair and using both signals?	D, E	ECE; fairness gap FG; reliance index MRI	Bootstrap; integrated gradients
RQ5	Is the gain worth the cost relative to the classical baseline?	B, E	Gain and compute relative to M0	Descriptive cost–benefit ranking

D. Measurement Model and Validity Criteria

The principal task metric M is macro-averaged F1 score for T1 and T3, giving equal weight to all classes, and the area under the precision–recall curve for T2, since arousal epochs constitute the minority class [36]. The complementarity measure CGt is the improvement of a fused model f relative to the better of the two single-signal models:

$$CGt = Mt(f) - \max(Mt(M1), Mt(M2)) \quad (1)$$

Retention with missing signal k is the ratio of the performance of the model on test with that signal removed to its performance when both signals are available, with the lowest value of the two cases given:

$$MMRm,k = M(m \mid k \text{ missing}) / M(m \mid \text{both present}), \; k \in \{ECG, EEG\}, \; MMRmin = mink \; MMRm,k \quad (2)$$

The cross-dataset drop, expected calibration error using 15 equal-sized bins [33], and the fairness gap across subgroups g are:

$$CDD = M(\text{source test}) - M(\text{target test}) \quad (3)$$

$$ECE = \Sigma b=115 \; (nb / n) \mid accb - confb \mid \quad (4)$$

$$FG = maxg \; M(g) - ming \; M(g) \quad (5)$$

There are six validity conditions which are determined prior to opening the test data. **V1**, no leakage: no subject appears in more than one of training, validation and test sets, determined through hash comparison of subject IDs; any overlap renders invalid the affected results. **V2**, adequate baselines: all models have the same tuning budget, and every deep single-signal model must be at least 20 percentage points above the majority-class baseline on the T1 validation set macro-F1 to compare the fused models with it. **V3**, gains beyond noise: a complementarity

The standard deviation of the metric of the model across the five seeds for training and its Holm-adjusted interval do not include zero. **V4**, matched capacity: the deep model should be within 20 percent of the parameter budget; otherwise, the gap will be reported. **V5**, alignment: those records where the ECG-EEG clock offset is more than 100 ms are discarded and reported. **V6**, reproducibility: the seeds, code, and configuration are frozen prior to evaluation, and the seed-to-seed variance is reported for every headline number.

Outcome measures- Table 6 lists the measures of outcomes, and Figure 7 shows the two key metrics.

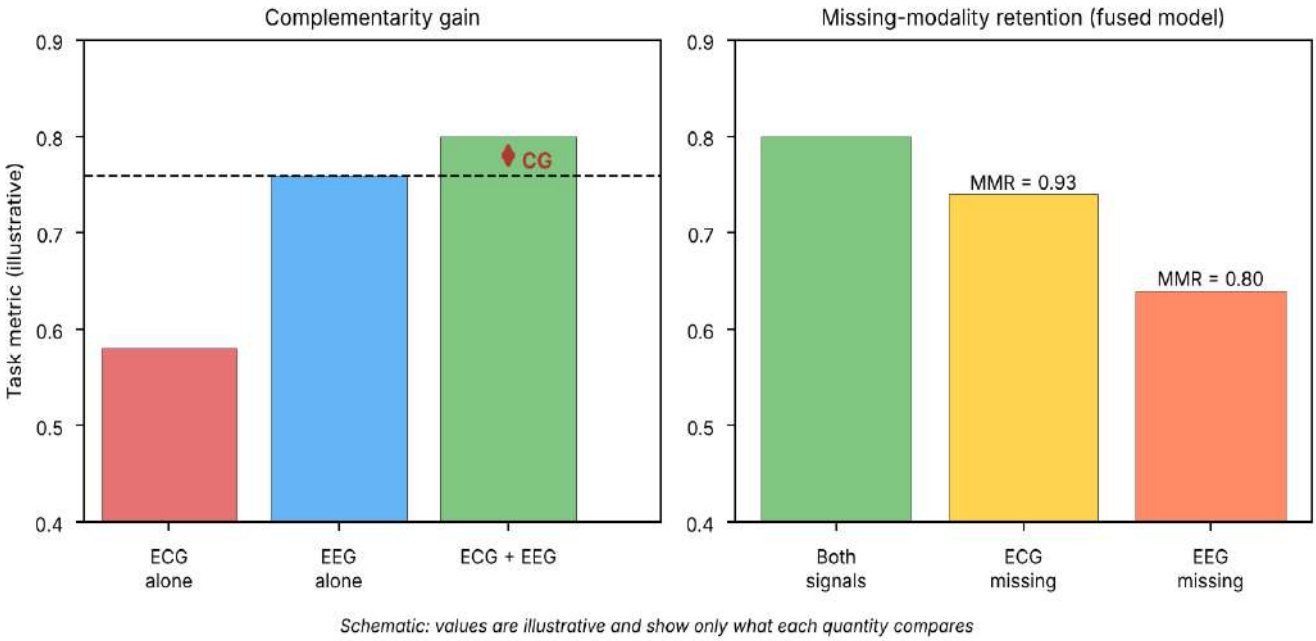

Figure 7: What the two central quantities compare. Left: complementarity gain over the best single-signal model. Right: retention when a signal is missing at test time. Values are illustrative

Table 6: Outcome measures

ID	Measure	Definition	Level
O1	Macro-F1	Mean F1 over classes (T1, T3)	Model by task
O2	Cohen's kappa	Agreement with scorers beyond chance [37]	Model, T1
O3	AUPRC and AUROC	Precision–recall and ROC areas for arousal [36], [38]	Model, T2
O4	CG	Complementarity gain, Eq. (1)	Fused model by task
O5	MMR	Missing-modality retention, Eq. (2)	Model by task
O6	CDD	Cross-dataset drop, Eq. (3)	Model by task
O7	ECE and Brier score	Calibration after temperature scaling on validation data [33], [39]	Model by task
O8	FG	Fairness gap across sex, age tertile and, where available, ethnicity, Eq. (5)	Model by task
O9	MRI	Share of attribution on ECG, Eq. (10)	Fused model
O10	Cost	Parameters, floating-point operations, latency and estimated energy	Model

IV. METHODOLOGY

The protocol includes the data acquisition and splitting (Phase A, Section 4.2), the signal processing and alignment (Section 4.3), the baselines, pre-training and training (Phases B and C, Section 4.4), the stress testing and transfer (Phase D, Section 4.5), the sample size and power calculations (Section 4.6), the statistical protocol (Phase E, Section 4.7), the schedule and computing (Section 4.8), the ethics and data governance (Section 4.9) and the threats to validity (Section 4.10). The protocol, the manifest files for the splits, the configurations and the scripts for the analyses are deposited in an open registry before any test data are released.

A. Programme Overview

The order of the steps is designed to protect test data. The splits are set up first, and test subjects are never part of any decisions. Hyperparameter selection, early stopping, and temperature scaling depend exclusively on validation data. Test sets are evaluated once the code and parameters are frozen. The study results are presented in line with guidelines for clinical prediction models built with machine learning techniques [35], with the intended use defined as research into signal fusion.

Prior to data access, Table 2 is re-coded by two independent reviewers in accordance with pre-defined search criteria for papers in PubMed, IEEE Xplore, ACM Digital Library, and arXiv, combining terms 'ECG', 'EEG' and 'multimodal', 'fusion'. Kappa statistic for inter-reviewer agreement [37] should be at least 0.61. In case the search results contradict the contents of the table, Table 2 is amended and this change is reported. Figure 8 shows the Architecture of the programme from data and splits.

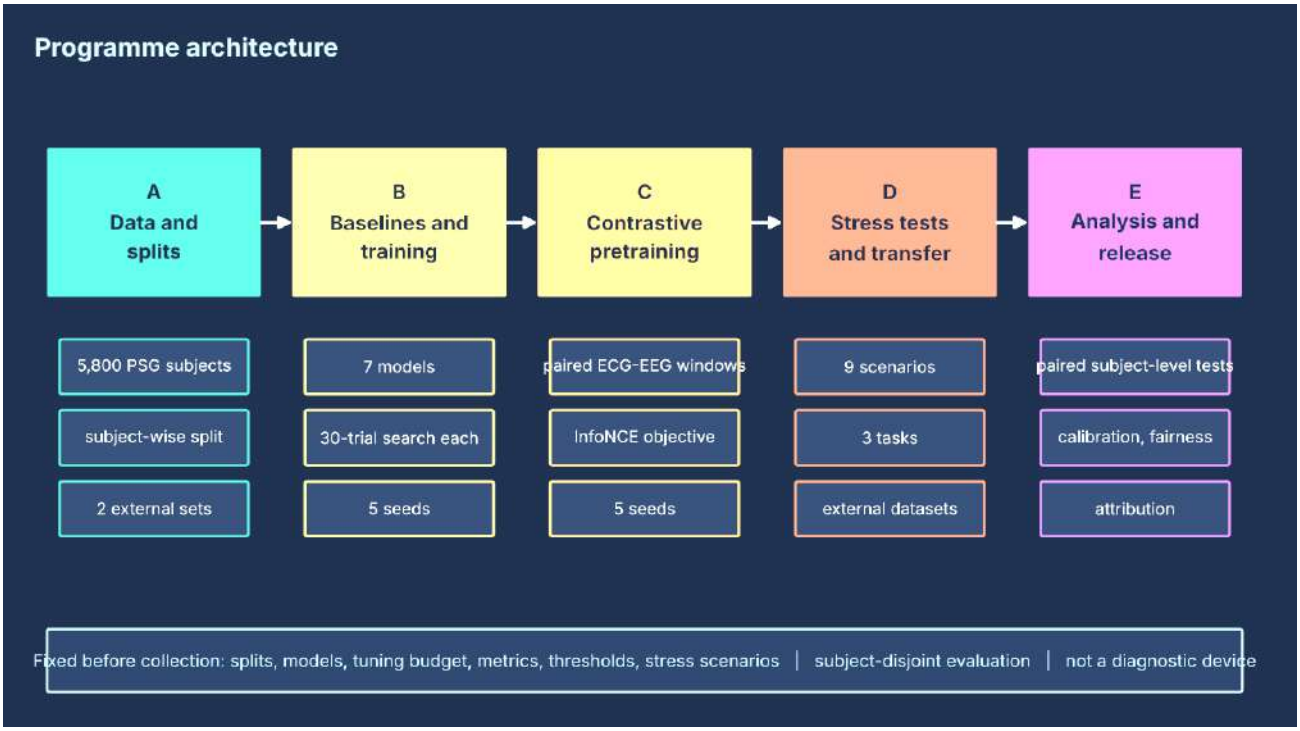

Figure 8: Architecture of the programme from data and splits through training and stress tests to analysis and release

B. Phase A: Data Access and Splits

D1 is a large population polysomnography cohort dataset available via a data use agreement from a sleep data common [25], [26], consisting of approximately 5,800 subjects having overnight ECG and central EEG, as well as sleep stage and arousal labels. This dataset is divided into training (3,480 subjects), validation (1,160 subjects) and test (1,160 subjects) datasets, stratified by sex and age range. D2 is the labelled subset of a polysomnography challenge set [27], [28], consisting of approximately 1,000 recordings which are never used in training or tuning but only as an external test set. D3 and D4 are affect datasets consisting of simultaneous EEG and ECG recordings during film watching [11], [12], comprising 23 and 40 subjects respectively. These are small datasets so they are tested via subject-independent 10-fold cross-validation performed three times with different random number generator seeds, and transfer between them by the shared channels and labels. All data set properties discussed in this paper are planning values, checked against the actual data dictionary at access time, and differences noted. See the figure 9 and Table 7.

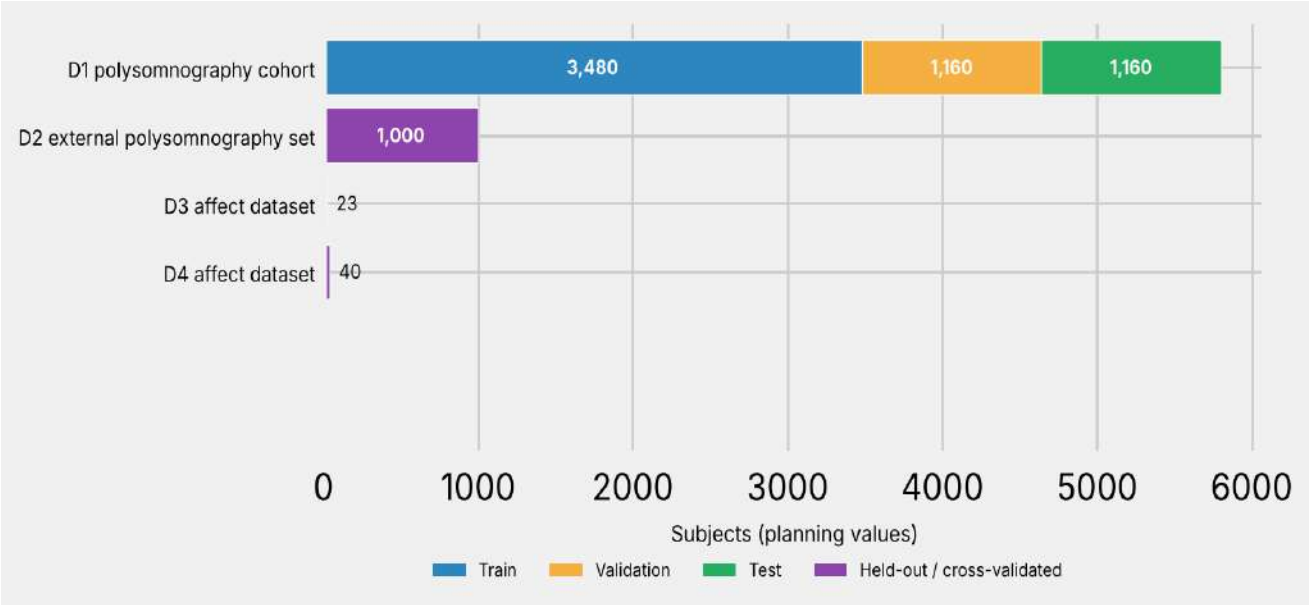

Figure 9: Planned allocation of subjects to training, validation, test and external or cross-validated roles in each dataset

Table 7: Data splits and stratification

Dataset	Subjects (planning)	Split	Stratification	Notes
D1 polysomnography cohort	5,800	3,480 train, 1,160 validation, 1,160 test (60, 20, 20 percent)	Sex and age band	Subject-disjoint, hashed identifiers (V1); test opened once
D2 external polysomnography set	about 1,000	External test only	None	Never used for training or tuning; montage mapped to D1
D3 affect dataset	23	10-fold subject-independent cross-validation, 3 seeds	Subject	Also transfer D3 to D4
D4 affect dataset	40	10-fold subject-independent cross-validation, 3 seeds	Subject	Also transfer D4 to D3; shared channels and labels only

C. Signal Processing and Alignment

Both the signals are separately processed through two pipelines until they come together during epoching (Table 8, Figure 10). ECG undergoes band-pass filtering, resampling to 128Hz and detection of R-peaks using a classical technique [40], which yields the heart-rate variability-based features used in M0 [17] along with the signal quality indicator. EEG undergoes band-pass filtering with a notch filter at mains frequency, resampling to 128Hz, re-referencing to the contralateral mastoid when possible and screening for gross artefacts. Artefactual epochs are not removed from the evaluation procedure: their labels are kept, the corresponding input is masked, and their proportion is counted to avoid biased evaluation against clean data.

Alignment of the two signals is verified before any modelling. Recordings are acquired with the same acquisition system and include event timestamps. The offset between the R-peaks and the artefact from cardiac fields observable in EEG is calculated by cross-correlation. Those recordings whose audited offset exceeds 100ms are discarded (V5), and their number is reported. Each epoch is a 30 second window, while each sample used for the model consists of the target epoch and two adjacent epochs on both sides (Table 9, Figure 11)._signals are standardised for each recording separately.

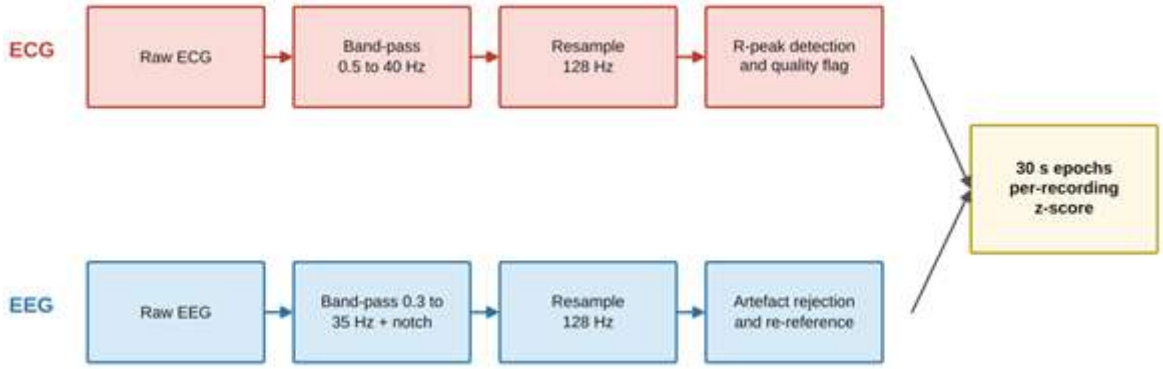

Figure 10: Processing pipelines for the two signals, which are independent until epoching

Table 8: Signal processing steps

Step	ECG	EEG	Rationale
Filtering	Band-pass 0.5 to 40 Hz	Band-pass 0.3 to 35 Hz plus mains notch	Removes drift and high-frequency noise
Resampling	128 Hz	128 Hz	Common rate across datasets and signals
Derived quantities	R-peaks [40]; heart-rate-variability features for M0 [17]	Band power in delta, theta, alpha, sigma and beta for M0	Inputs to the classical baseline
Quality control	Flag epochs with implausible beat counts or flat lines	Flag epochs with extreme amplitude	Flagged epochs are masked, not dropped
Re-referencing	Single lead	Contralateral mastoid where available	Comparable montages across D1 and D2
Normalisation	Per-recording z-score	Per-recording z-score	Reduces amplitude differences between devices

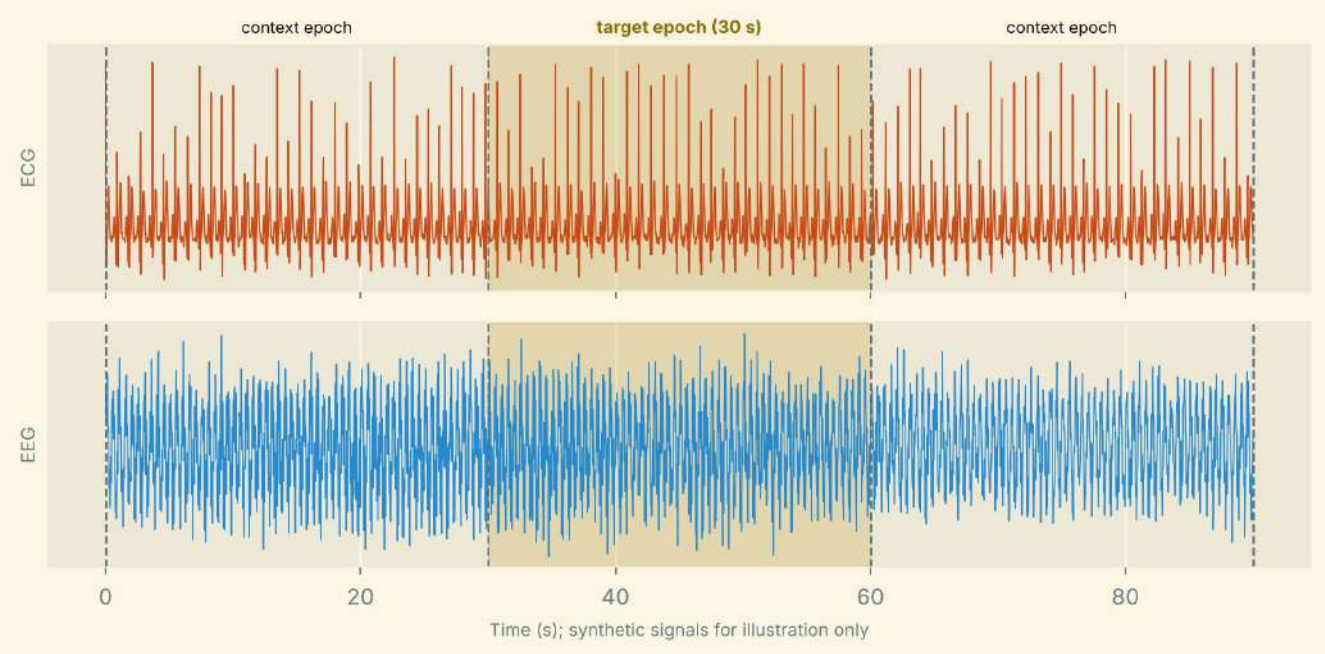

Figure 11: Windowing used throughout: a 30 second target epoch with two context epochs on each side. The signals shown are synthetic and for illustration only

Table 9: Windowing and alignment parameters

Parameter	Value	Note
Common sampling rate	128 Hz	3,840 samples per 30 s epoch
Target epoch	30 s	Matches the scoring unit
Context	Two epochs before and two after	Five-epoch sequence per sample
ECG channels	One lead	Lead I or the nearest available
EEG channels	Two central channels (D1, D2); 14 channels (D3, D4)	Montage mapped between D1 and D2; shared channels between D3 and D4
Clock-offset limit	100 ms	Recordings above the limit are excluded (V5)
Artefact handling	Mask input, keep label	Reported as a proportion
Augmentation	Time shift up to 2 s; amplitude scaling 0.9 to 1.1; additive noise	Training only

D. Phases B and C: Baselines, Pretraining and Training

Training protocol- All deep models are fine-tuned using the same optimizer, scheduler, and budget (Table 10, Figure 12): AdamW [41] with cosine decay following a two-epochs warm-up, batch size 128 for five-epoch sequences up to 30 epochs based on early stopping on the validation metric, mixed precision training, class-weighted cross-entropy for handling class imbalance. Each architecture is fine-tuned with 30 trials of random search on the T1 validation set, using the same search space for shared hyperparameters, and the best-chosen architecture is used again for T2 and T3 with the learning rate and weight decay being re-tuned for 10 trials. Each trial is trained with 5 different seeds. The temperature scaling is fit to the validation set after the training [33].

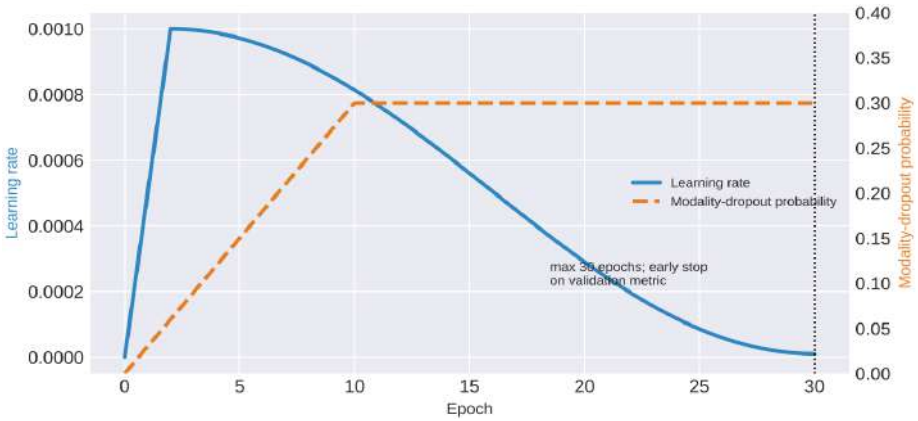

Figure 12: Learning-rate schedule (warm-up then cosine decay) and the ramp of the modality-dropout probability over the first ten epochs

Table 10: Training configuration

Setting	Value	Applies to
Optimiser	AdamW, weight decay 0.01 [41]	All deep models
Learning rate	Peak 1e-3, two-epoch warm-up, cosine decay to 1e-5	All deep models
Batch and epochs	128 sequences; up to 30 epochs; early stopping	All deep models
Loss	Class-weighted cross-entropy	All deep models
Seeds	5 for T1 and T2; 3 for T3	All models
Tuning budget	30 random-search trials on the T1 validation set; 10 trials for T2 and T3 re-tuning	All models including M0
Parameter budget	Within 20 percent of 2 million parameters	Deep models (V4)
Modality dropout	Probability ramped to 0.30 over 10 epochs; at least one signal always present	M5 and M6 (M3 and M4 as ablations)
Calibration	Temperature scaling fitted on validation data	All models

Contrastive pretraining of the bridge model- M6 is identical to M5 except for its initialisation and training. Both of its encoders are pretrained in the unsupervised manner on co-registered windows of ECG and EEG of D1 subjects without labels using the contrastive learning objective. Namely, embeddings of the same 30-second window of ECG and EEG are made close to each other while embeddings of other windows of the batch are made distant [31], [32](Figure 13 is showing the schematic of the contrastive objective for a batch of eight windows). Given embeddings zh and zb of heart and brain windows, batch size B and temperature τ, the loss (in one direction and symmetrised) is

$$L_{NCE} = -(1/B)\ \Sigma i \log [\ \exp(\ sim(zh,i,\ zb,i)\ /\ \tau\ )\ /\ \Sigma j \exp(\ sim(zh,i,\ zb,j)\ /\ \tau\ )\ ] \quad (6)$$

In the negative examples, we used windows from different moments of time of the same person, so that the model had to align the signals within the night and not just learn the identity of the person. After the unsupervised pretraining, the encoders are merged through the cross-attention module, where queries are taken from one signal, while the keys and values are taken from the other [30]:

$$Attn(Q, K, V) = softmax(\ Q\ KT\ /\ \sqrt{d}\ )\ V \quad (7)$$

After that, the whole model was fine-tuned in the supervised manner with additional contrastive loss and modality dropout. Modality dropout is the probabilistic dropping out of either ECG or EEG presence masks mh and mb, each with probability 1 − p but at least one of them [13]:

$$L = L_{CE} + \lambda\ L_{NCE},\ \ \tilde{x} = (\ mh \odot xh,\ mb \odot xb\ ),\ \ m \sim Bernoulli(1 - p),\ \ mh + mb \geq 1 \quad (8)$$

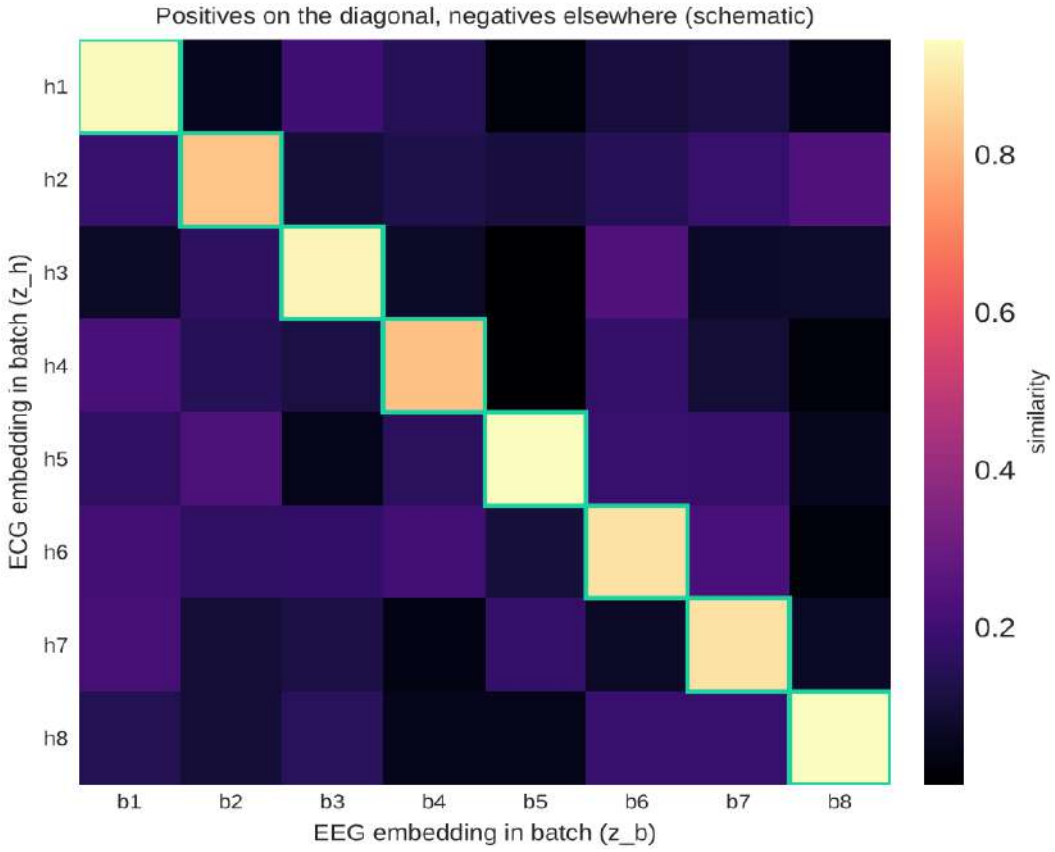

Figure 13: Schematic of the contrastive objective for a batch of eight windows: matched ECG and EEG windows (diagonal) are positives and all other pairs are negatives

Table 11: Contrastive pretraining and bridge fine-tuning

Setting	Value	Note
Pretraining data	Co-registered ECG and EEG windows from D1 training subjects	No labels; no validation or test subject
Objective	Symmetric InfoNCE, Eq. (6)	[31]
Temperature τ	0.07	Fixed; not tuned on test data
Batch size	256 windows	Negatives from the batch, including same-subject other times
Projection head	128-dimensional	Discarded after pretraining
Pretraining epochs	50	Five seeds
Fine-tuning	Supervised loss plus λ = 0.5 contrastive term, modality dropout p = 0.30	Eq. (8)
Ablations	M5 (no pretraining), M6 without modality dropout, M6 without pretraining	Isolate each component

E. Phase D: Stress Tests and Transfer

Every trained network is then tested on the test datasets with respect to nine scenarios (Table 12, Figure 14). The missing signal is modeled by replacing the signal with zero values while setting the corresponding presence mask; the missing feature block for M0 is replaced with the mean values from the training set. For noise scenarios, a white Gaussian noise is added to a single signal with respect to a fixed signal-to-noise ratio calculated per window. The channel-drop scenario is modeled by assigning zero values to half of the EEG channels randomly. Cross-dataset testing consists of applying the trained networks to D1 on D2 and vice versa, and D3 on D4 and vice versa. All scenarios are run only on the test data, with the networks and threshold values held constant.

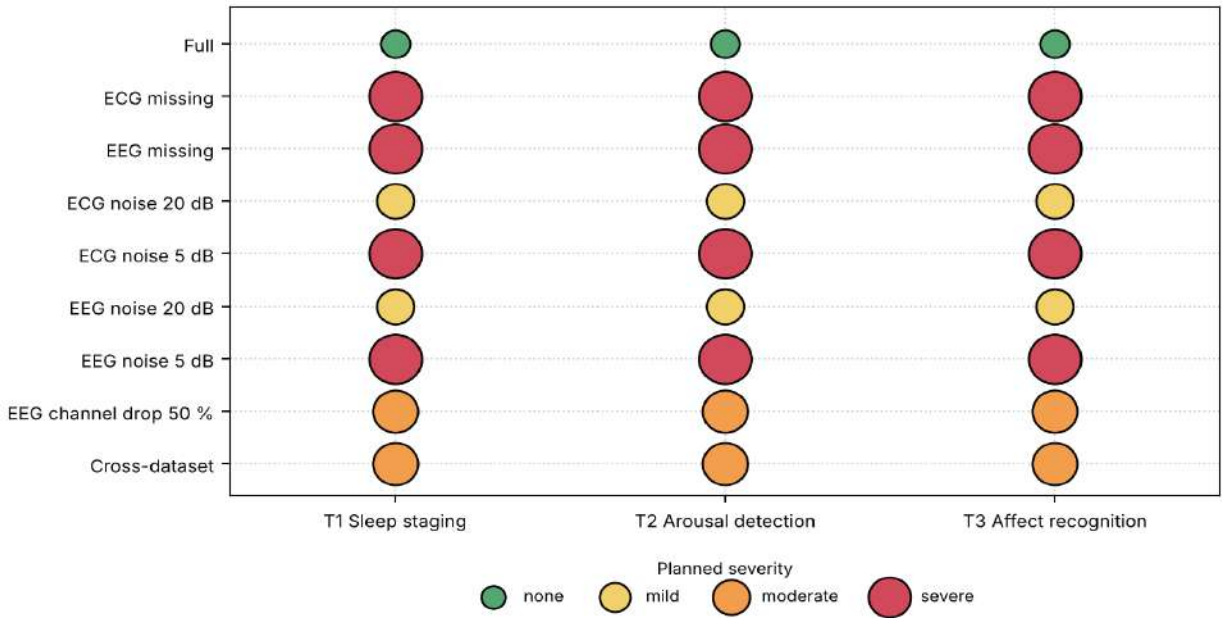

Figure 14: Planned stress-test grid: each of nine scenarios is applied to each of the three tasks, with the severity indicated by colour and size

Table 12: Stress-test scenarios.

Scenario	Implementation	Severity	Applies to
Full	Both signals present	None	All tasks
ECG missing	ECG replaced by zeros and mask set (M0: mean features)	Severe	All tasks
EEG missing	EEG replaced by zeros and mask set (M0: mean features)	Severe	All tasks

ECG noise 20 dB	White noise at 20 dB signal-to-noise ratio	Mild	All tasks
ECG noise 5 dB	White noise at 5 dB signal-to-noise ratio	Severe	All tasks
EEG noise 20 dB	White noise at 20 dB signal-to-noise ratio	Mild	All tasks
EEG noise 5 dB	White noise at 5 dB signal-to-noise ratio	Severe	All tasks
EEG channel drop 50 %	Half of the EEG channels zeroed at random	Moderate	All tasks
Cross-dataset	Train on source, test on target: D1 to D2; D3 to D4 and D4 to D3	Moderate	T1 and T2 (D1 to D2); T3 (D3, D4)

F. Sample Size and Power

The confirmatory contrasts are made between paired models at the level of the individual. Each individual and each task, a comparison is made between the metric of the fused model (M5 or M6) and the metric of the best single-signal model, and the unit of analysis is the per-individual difference ds. For a paired test of a mean difference δ with standard deviation σd of the differences at significance α and power 1 − β, the sample size n is [42]:

$$n \approx [(z_{1-\alpha/2} + z_{1-\beta}) \sigma_d / \delta]^2, \quad \alpha = 0.05 / 6 \quad (9)$$

The six confirmatory contrasts (three tasks, each for M5 and M6 against the best single-signal model) have a familywise error rate of 0.05 and are thus tested at α = 0.0083. With a smallest difference of interest δ = 0.02 in the metric and an assumed σd = 0.06 and power 0.80, the exact calculation based on the noncentral t distribution yields n = 113 test individuals. If δ were 0.01, n = 439 would be needed, and if δ were 0.03, n = 52 (Figure 15). If σd were 0.10, then the requirement for δ = 0.02 would increase to 306. The design thus provides adequate power for T1 and T2: the D1 test set of 1,160 individuals can detect a difference of about 0.006, and the D2 external set of about 1,000 individuals about 0.007, with 80 percent power. For T3, the 63 individuals of the two affect datasets, with an assumed σd of 0.08 due to greater individual-to-individual variability, can detect a difference of about 0.036; smaller differences are treated as exploratory. Since gains in the order of magnitude of seed noise cannot be seen as evidence, criterion V3 requires that a gain be more than twice the standard deviation of the gains from one seed to another.

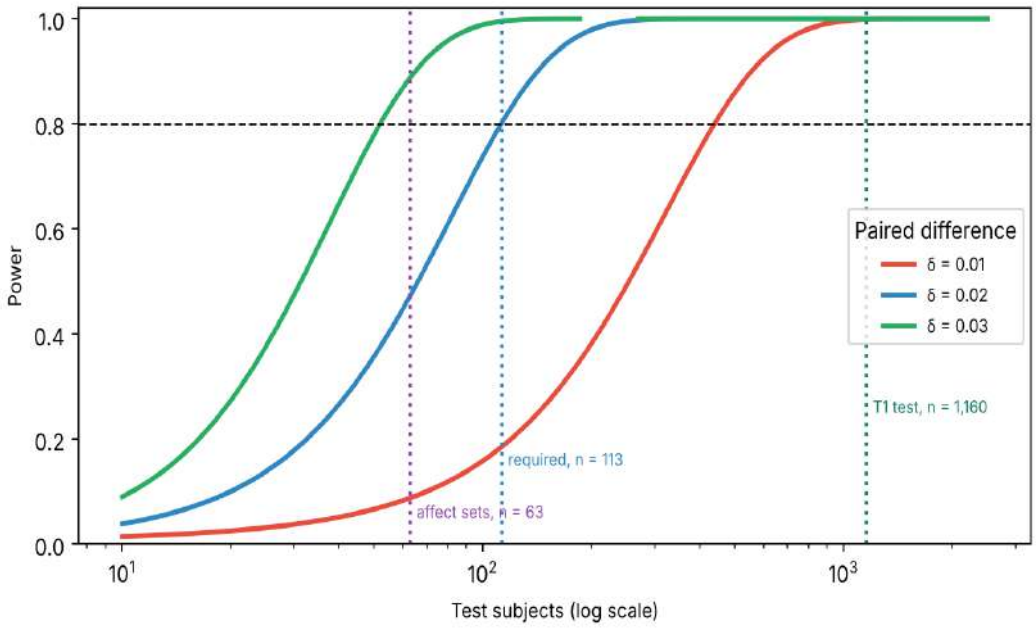

Figure 15: Power of the paired test as a function of the number of test subjects for three paired differences, with the sizes of the available test sets marked (α = 0.05 / 6, assumed σ_d = 0.06)

Table 13: Sample size and power

Parameter	Value	Basis
Unit of analysis	Test subject; per-subject metric averaged over seeds	Avoids pseudo-replication
Confirmatory contrasts	6 (3 tasks x M5, M6 against best unimodal)	Family-wise α = 0.05
Adjusted significance level	0.0083	Bonferroni; Holm for follow-up [43]
Smallest difference of	0.02	Planning assumption

interest δ		
SD of paired differences σ_d	0.06 (T1, T2); 0.08 (T3)	Planning assumption; sensitivity above
Power	0.80	Convention
Required n	113	Eq. (9), noncentral t
D1 test subjects	1,160	Minimum detectable difference about 0.006
D2 external subjects	about 1,000	Minimum detectable difference about 0.007
Affect participants	63	Minimum detectable difference about 0.036; exploratory below this
Seeds	5 (T1, T2); 3 (T3)	Seed SD enters criterion V3

G. Phase E: Statistical Analysis Protocol

All analyses are pre-specified prior to unblinding of the testing data. The main comparison within each task is the per-subject difference between a fused model and the best performing single-signal model, using a paired t-test in case of passing normality tests of differences and Wilcoxon signed-rank test otherwise [44], with Holm correction for multiple comparisons across the six confirmatory comparisons [43]. Confidence intervals are bias-corrected and accelerated bootstrap intervals with re-sampling of subjects, 10,000 repeats [45]. The seven models are ranked by the per-subject measure using Friedman test, followed by post-hoc Wilcoxon contrast tests with Holm correction, because of the issues of multiplicity and dependence when evaluating [46]. Point estimates of classifier performance without confidence intervals are not reported. MMR, CDD (Eqs. 2, 3) and their bootstrap confidence intervals summarize robustness. ECE (Eq. 4) and Brier score [39] before and after temperature scaling are used to evaluate calibration. Subgroup analysis includes performance estimation for sex, age tertiles and ethnicity, where available, with fairness gap (Eq. 5) and bootstrap confidence interval. Since subgroups can be quite small, gaps are presented along with their sample size and not tested against zero. A gap of 0.02 or higher in the primary performance measure is pre-defined as practically important [34]. Modality reliance is assessed through integrated gradients [47] calculated on 500 testing samples per class, and the reliance index is

$$MRI = (\Sigma | IGECG |) / (\Sigma | IGECG | + \Sigma | IGEEG |) \quad (10)$$

The model ignoring one of the signals would have MRI equal to 0 or 1, and the MRI is considered in conjunction with the missing-modality results. Missing predictions are not imputed, if a recording is missed according to V5 criteria, then it is missed for all models. Figure 16 and Table 14 show decision rules.

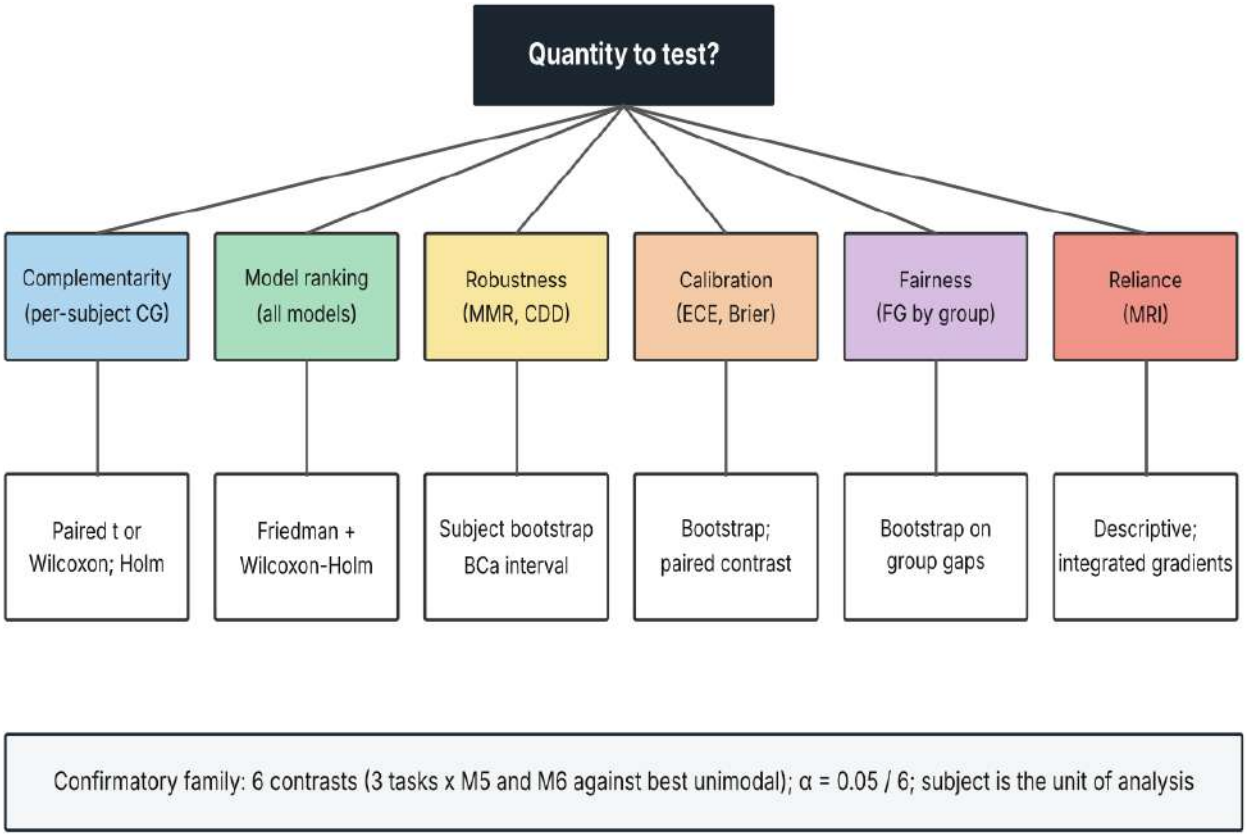

Figure 16: Decision tree for the choice of analysis by quantity. The confirmatory family is the six contrasts of M5 and M6 against the best single-signal model in each task

Table 14: Statistical analysis plan and decision rules

Quantity	Primary analysis	Assumption check	Decision rule
Complementarity (CG)	Paired t or Wilcoxon on per-subject differences, Eq. (1)	Normality of differences	V3: exceeds 2 seed SD; Holm-adjusted interval excludes 0
Model ranking	Friedman test; Wilcoxon contrasts [46]	Paired subjects	Holm-adjusted
Robustness (MMR, CDD)	Subject bootstrap, BCa [45]	Resample subjects	MMR threshold 0.85 for "robust" (Table 21)
Calibration	ECE and Brier before and after scaling [33], [39]	15 equal-width bins	Report; H5 compares models
Fairness	Group gaps with bootstrap intervals, Eq. (5)	Group sizes reported	Gap of 0.02 or more is practically important
Reliance	Integrated gradients, Eq. (10) [47]	500 stratified test epochs	Descriptive
Seed variability	SD across seeds per model	5 seeds (T1, T2); 3 (T3)	V3 threshold
Cost	Parameters, operations, latency, energy	Same hardware	Descriptive ranking against M0

H. Schedule, Compute and Reproducibility

The program has a 20-week duration and will require approximately 619 GPU-hours (Table 15 and Figure 17). With an assumption that the average draw is 0.3 kW, the power consumption is estimated at 186 kWh, which is reported alongside the results as the environmental cost. The hyperparameter search is the most significant cost component, while the pre-training stage of the bridge model, even though it is relatively expensive on an individual run basis, represents only a minor part of the total cost. All random numbers are generated based on the seed and saved with the results.

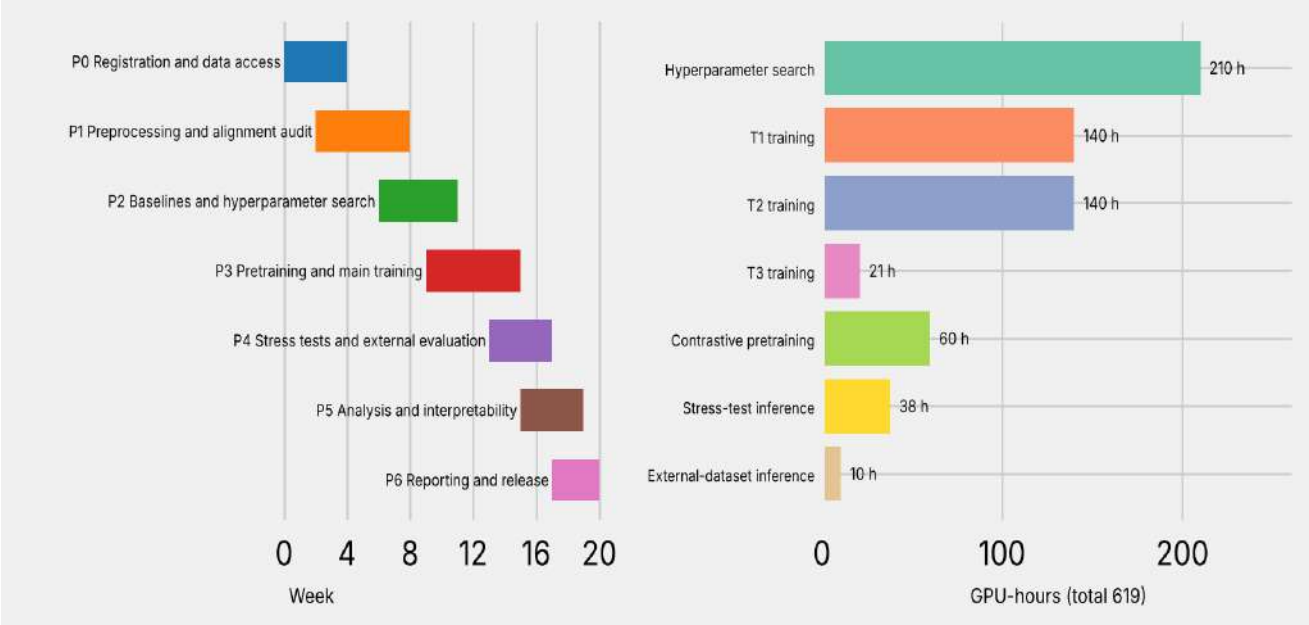

Figure 17: Left: schedule of the seven phases over 20 weeks. Right: planned GPU-hours by activity

Table 15: Schedule and compute budget

Activity	Runs	GPU-hours each	GPU-hours
Hyperparameter search (7 models, 30 trials)	210	1	210
T1 training (7 models x 5 seeds)	35	4	140
T2 training (7 models x 5 seeds)	35	4	140
T3 training (7 models x 10 folds x 3 seeds)	210	0.1	21
Contrastive pretraining (5 seeds)	5	12	60
Stress-test inference (7 models x 9 scenarios x 3 tasks)	189	0.2	38
External-dataset inference	21	0.5	10
Total			619

I. Ethics, Governance and Intended Use

No recruitment of participants occurs in this study, as existing datasets are used. Polysomnography dataset access is under the data-use agreement with the data commons, and other datasets are accessed under their licenses. No attempts to redistribute data for purposes different from the specified, no attempts to re-identify participants, and no redistribution of raw signals are performed. Physiological signals are identifying information: an ECG contains personal data of the subject, and the pretrained weights may contain subject-specific information. Thus, the protocol keeps raw signals and extracted features in secure repositories, releases the weights only after verification that the identification of the subject from embeddings is not possible above the chance level, and uses hashed subject identifiers in the split manifests. Personal data, when protected by data protection laws, are managed according to those requirements [48].

Intended use of the models is research into signal fusion. Models are not diagnostic tools, no output is presented as a recommendation for clinical decision-making, and any use of the models in clinical context will require validation. Model performance in subgroups, calibration properties and behaviour in case of missing modalities are reported because they affect safety of using the model. Reporting is done following the recent guidance for prediction models based on machine learning [35]. Release levels of each artefact are shown in Table 16 and Figure 18.

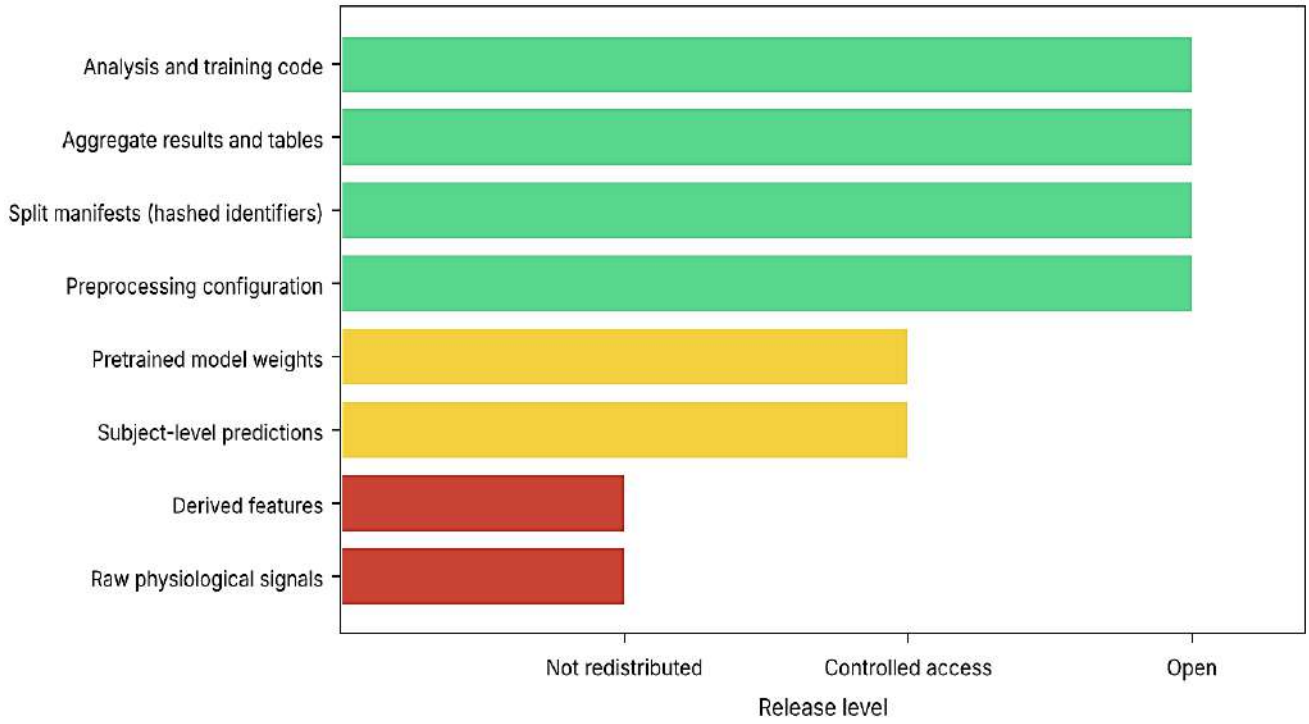

Figure 18: Release level of each artefact: open, controlled access or not redistributed

Table 16: Artefacts, release levels and rationale

Artefact	Release level	Rationale
Analysis and training code	Open	Supports replication; contains no data
Aggregate results and tables	Open	Supports scrutiny of conclusions
Split manifests (hashed identifiers)	Open	Lets others reproduce the splits without exposing identities
Preprocessing configuration	Open	Needed to reproduce the inputs
Pretrained model weights	Controlled access	Released after a subject-identity recoverability check
Subject-level predictions	Controlled access	Could aid re-identification; controlled access with stated purpose
Derived features	Not redistributed	May identify individuals; follows source data-use terms
Raw physiological signals	Not redistributed	Governed by the data-use agreements; never redistributed

J. Threats to the Validity of the Programme

Such a program, in turn, is also an instrument. Here, ten threats are mentioned ahead of time, and Table 17 shows the threat together with its mitigation. They are placed into the likelihood-impact matrix in Figure 19.

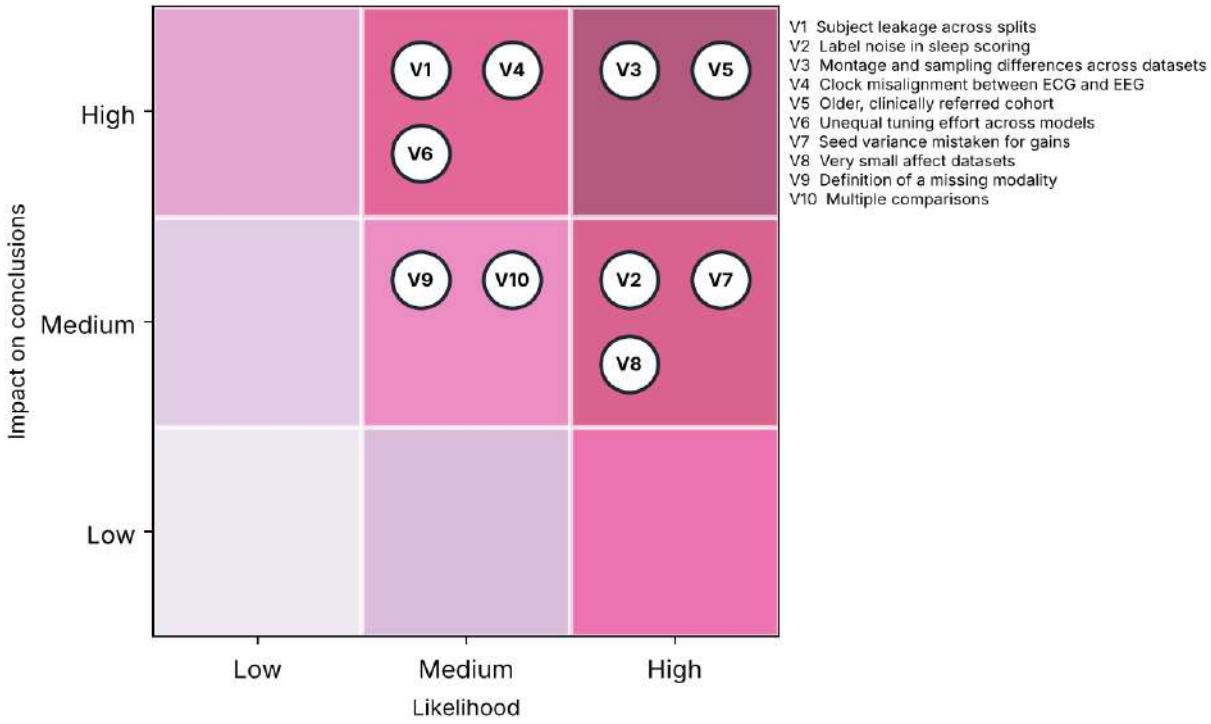

Figure 19: Likelihood and impact of the ten named threats to validity, assessed before collection. (Identifiers refer to Table 17)

Table 17: Threats to validity and mitigation

ID	Threat	Likelihood	Impact	Mitigation
V1	Subject leakage across splits	Medium	High	Subject-wise splits; hashed-identifier check (criterion V1); test data opened once
V2	Label noise in sleep scoring	High	Medium	Report scorer agreement; use large samples; interpret small differences cautiously
V3	Montage and sampling differences across datasets	High	High	Map montages; report D1 to D2 transfer; stress-test channel loss
V4	Clock misalignment between ECG and EEG	Medium	High	Audit offsets; exclude recordings above 100 ms (criterion V5); report exclusions
V5	Older, clinically referred cohort	High	High	State the population; stratify by age and sex; report subgroup results
V6	Unequal tuning effort across models	Medium	High	Equal tuning budget for all models, including M0; parameter budget within 20 percent
V7	Seed variance mistaken for gains	High	Medium	Five seeds; require a gain above twice the seed SD (criterion V3)
V8	Very small affect datasets	High	Medium	Treat T3 as exploratory below the minimum detectable difference; report intervals
V9	Definition of a missing modality	Medium	Medium	Define missing as zeros plus mask for deep models and mean features for M0; report both definitions
V10	Multiple comparisons	Medium	Medium	Holm adjustment across the six confirmatory contrasts

V. PRE-REGISTERED PREDICTIONS AND REPORTING STRUCTURE

In this part, the analysis is defined before data is obtained. The number of items to be reported is defined and the predicted direction of the results is also defined along with the rationale behind each prediction. The current paper does not have an estimate as the process detailed in Section 4 has not been carried out yet. Predictions are formed based on the literature and based on how each particular model works and predictions in the program are expected to be wrong. Tables 18 and Figure 20 contain the predictions, while Tables 19 and Figure 21 contain the reporting format for models. Figure 22 shows the template for the retention of each model

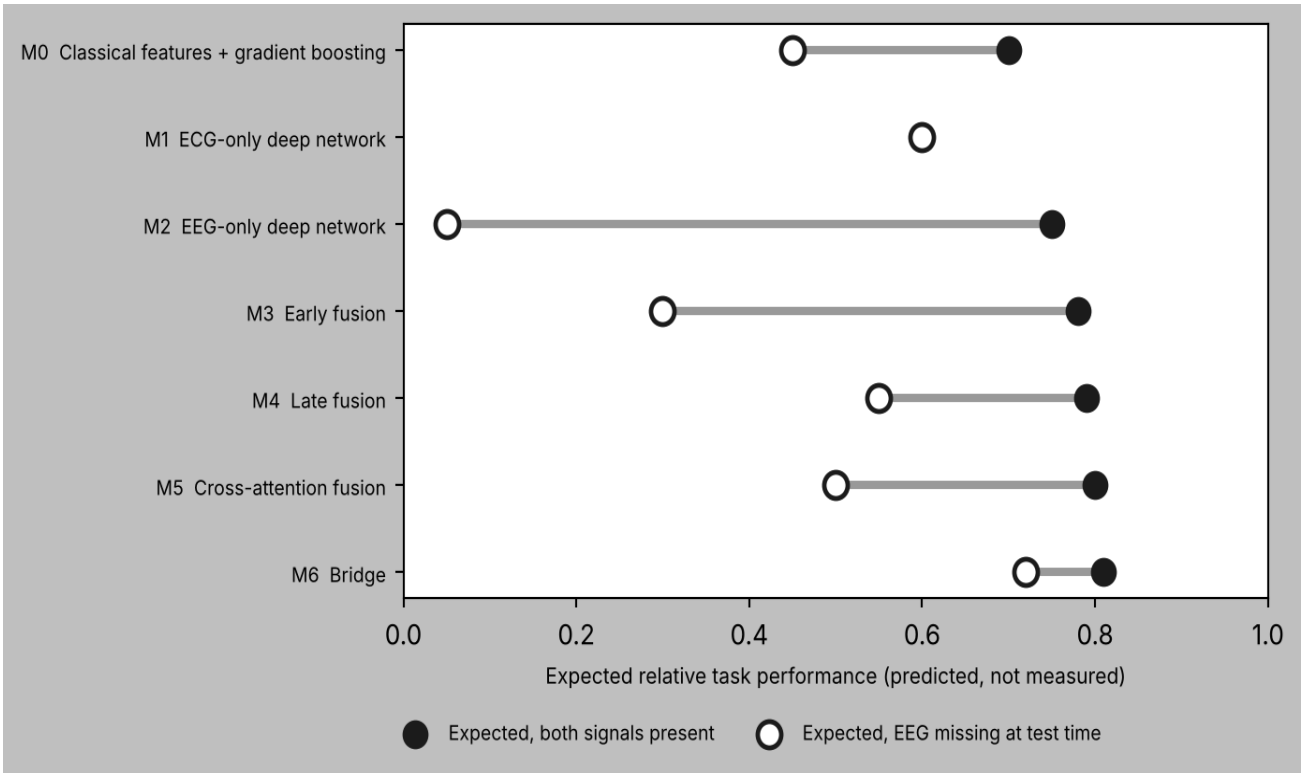

Figure 20: Expected relative performance of each model with both signals present and with EEG missing at test time. Positions are predictions recorded before collection and are not measured values

Table 18: Pre-registered hypotheses

ID	Hypothesis	Predicted direction	Basis	Test
H1	Complementarity depends on the task	CG for arousal detection larger than for sleep staging; CG for staging at most 0.03	Arousals are cortical events with autonomic responses; EEG dominates staging [24], [8]	Paired test, Eq. (1)
H2	EEG alone outperforms ECG alone in staging	Macro-F1 of M2 exceeds M1 by at least 0.10	Staging labels are defined on EEG [23], [10]	Paired test
H3	Missing-modality behaviour differs by fusion	MMR with EEG missing at least 0.85 for M6 and below 0.70 for M3	Modality dropout supports robustness [13]	Subject bootstrap
H4	Fusion reduces the cross-dataset drop	CDD of the best fused model no larger than that of the best unimodal model	Complementary signals may offset a shift in one	Subject bootstrap
H5	Fused models are less calibrated before scaling	ECE of fused models higher than unimodal before scaling; at most 0.05 after	Modern networks are overconfident [33]	Bootstrap
H6	Subgroup gaps persist under fusion	FG across age tertiles of at least 0.02 and not smaller for fused models	Age-related signal differences [34]	Bootstrap
H7	A strong classical baseline is competitive on small data	M0 within 0.05 of the best deep model on T3 but not on T1	Few subjects favour hand-crafted features [22]	Paired test
H8	Fused models use both signals	MRI between 0.2 and 0.5 for M5 and M6 on T2	Both signals carry arousal information	Integrated gradients

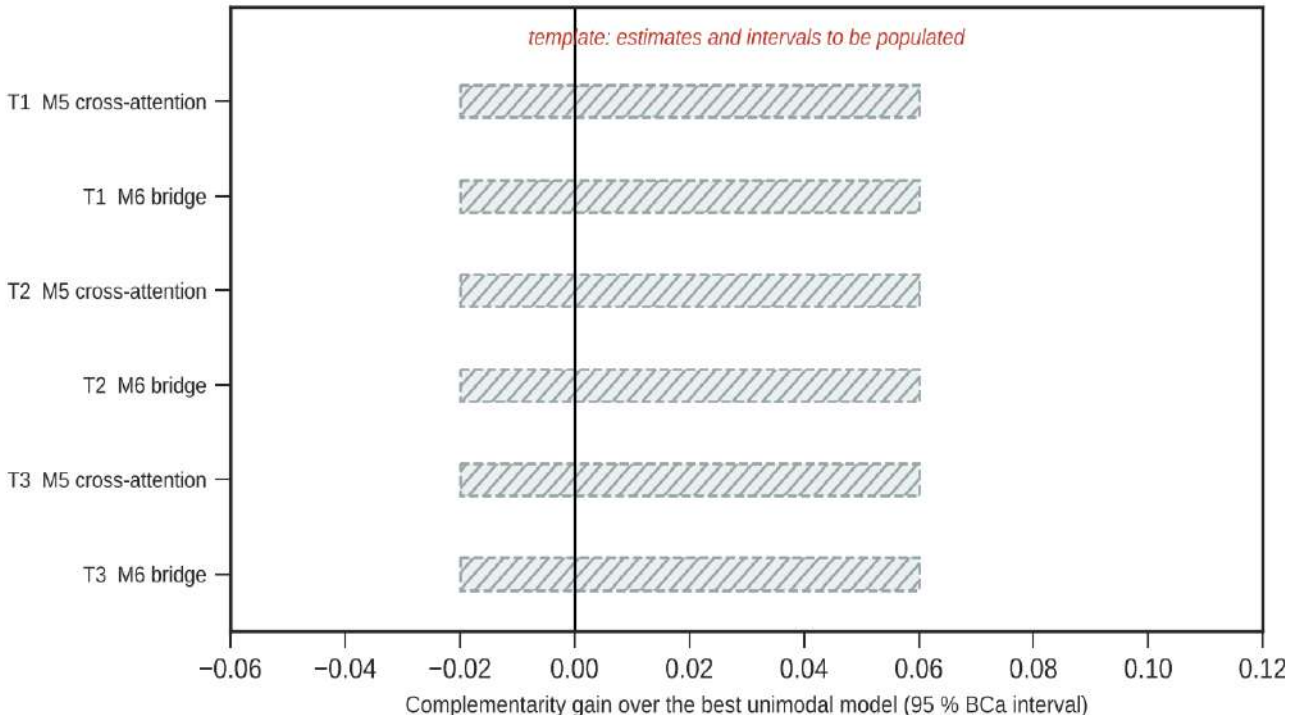

Figure 21: Template for the complementarity gain of M5 and M6 over the best single-signal model on each task. The hatched bars mark where each estimate and its 95 percent interval will be placed

Table 19: Reporting template by model. To be populated from the collection of Section 4

Model	T1 macro-F1	T1 kappa	T2 AUPRC	T3 macro-F1	ECE	CG (T1, T2, T3)	MMR min	Parameters
M0 Classical features + gradient boosting	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M1 ECG-only deep network	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M2 EEG-only deep network	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M3 Early fusion	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M4 Late fusion	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M5 Cross-attention fusion	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M6 Bridge	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate

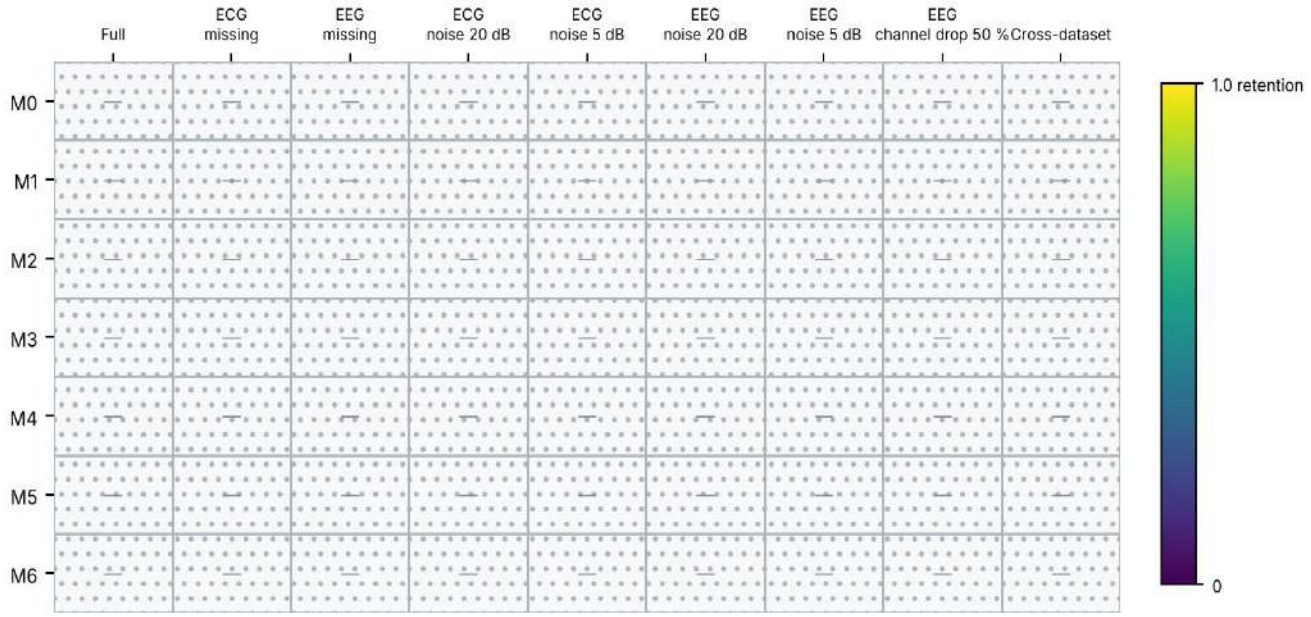

Figure 22: Template for the retention of each model under each stress scenario. Cells will be shaded by retention from 0 to 1

Table 20: Reporting template for stress tests. To be populated from the collection of Section 4

Model	ECG missing	EEG missing	ECG noise 5 dB	EEG noise 5 dB	EEG channel drop	Cross-dataset drop	Subgroup gap (age)
M0 Classical features + gradient boosting	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M1 ECG-only deep network	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M2 EEG-only deep network	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M3 Early fusion	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M4 Late fusion	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M5 Cross-attention fusion	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate
M6 Bridge	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate	To estimate

VI. DISCUSSION

Interpretation of each outcome is pre-specified such that discussion is not written to suit the data shown. The Table 21 shows six possible outcomes, while Figure 23 describes complementarity relative to robustness. If the performance of a fused model exceeds the best performing single-signal model by more than seed noise and retains at least 85% performance if any signal is lost, then the use of two signals is justified in the context of the task performed, and any additional mechanism in the bridge model will be justified if it performs better than simple fusion models by the same standard. This result still requires further investigation and not clinical use of the technology because it will require prospective validation in the target population. If the fused model exhibits complementarity but lack robustness, the use of two sensors will provide added accuracy if and only if both signals are present. In this case, it reduces the utility of the model in settings with missing signals. If there is no complementarity between the two signals, one of the signals cannot be justified since the extra signal does not compensate for the added accuracy of the best single-signal model.

There are three additional results that are relevant regardless of the primary outcome. If the classical baseline performs within 0.05 of the deep models, there is little justification of deep learning in that task. Also, if subgroup disparities do not reduce or increase with fusion, there is the danger of perpetuating disparities with the improvement in the overall performance with fusion. Finally, if all the models have a huge cross-dataset drop, then the results from the single cohort cannot be extrapolated and reported as such. It should be noted that a negative result does not detract from the efforts used to build the programme since physiologically coupled signals are not necessarily complementary (See the Figure 24).

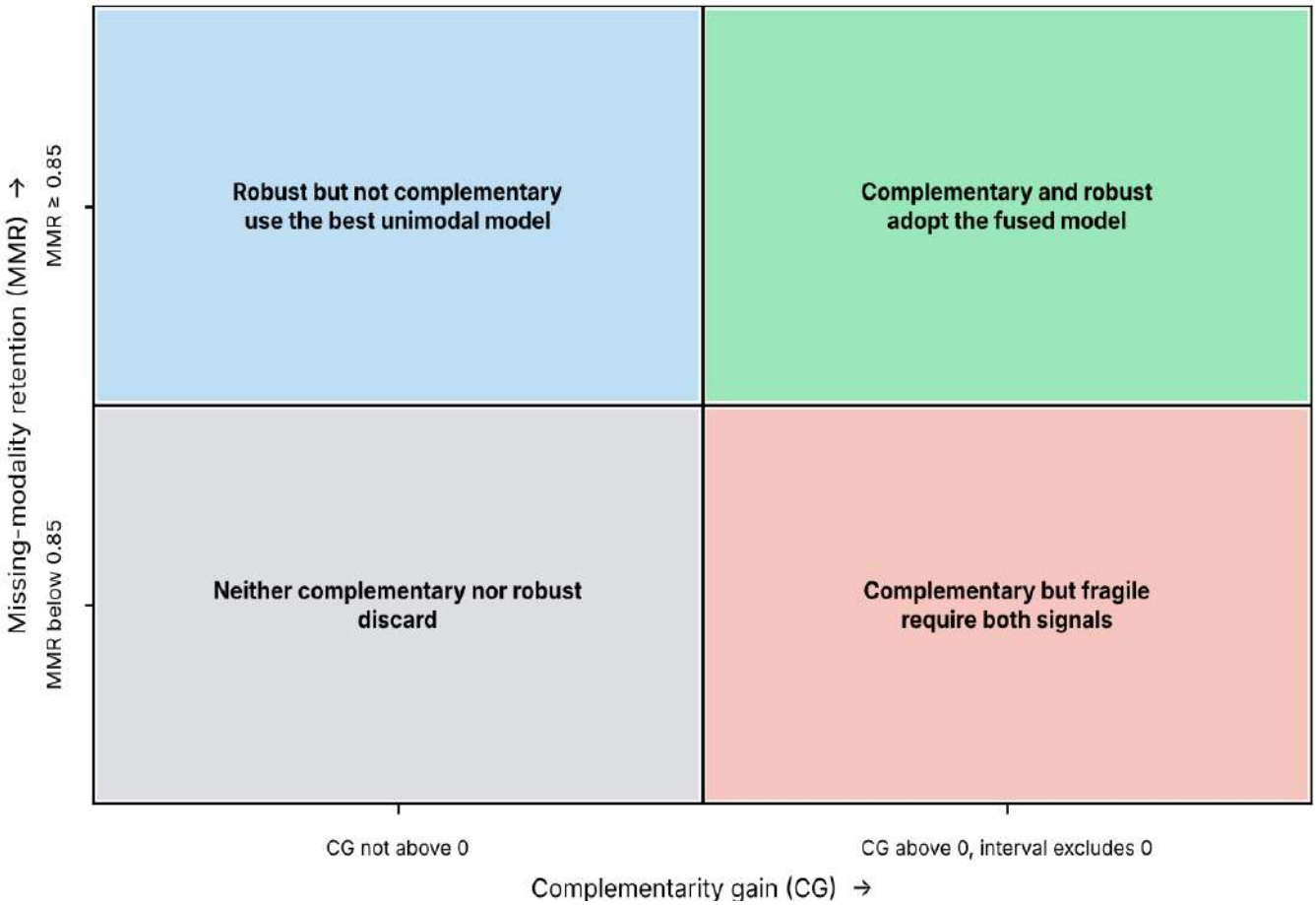

Figure 23: Interpretation map. Each fused model will be placed in one of four regions by its complementarity gain and its missing-modality retention

Table 21: Interpretation fixed in advance for each pattern of outcomes

ID	Observed pattern	Interpretation	Implication
O1	CG above 0 by criterion V3 and MMR at least 0.85	Complementary and robust	Using both signals is justified for the task; further validation needed before any clinical claim
O2	CG above 0 but MMR below 0.85	Complementary but fragile	Both signals required; unsuitable where a signal is often missing
O3	CG not above 0	No complementarity	Use the best single-signal model; the second sensor adds no accuracy
O4	M0 within 0.05 of the best deep model	Deep learning not justified on this task	Prefer the simpler model
O5	Subgroup gap not smaller under fusion	Fusion does not reduce disparities	Report and investigate before use
O6	Large CDD for all models	Results do not transfer across cohorts	Single-cohort claims are unsupported

Nomenclature

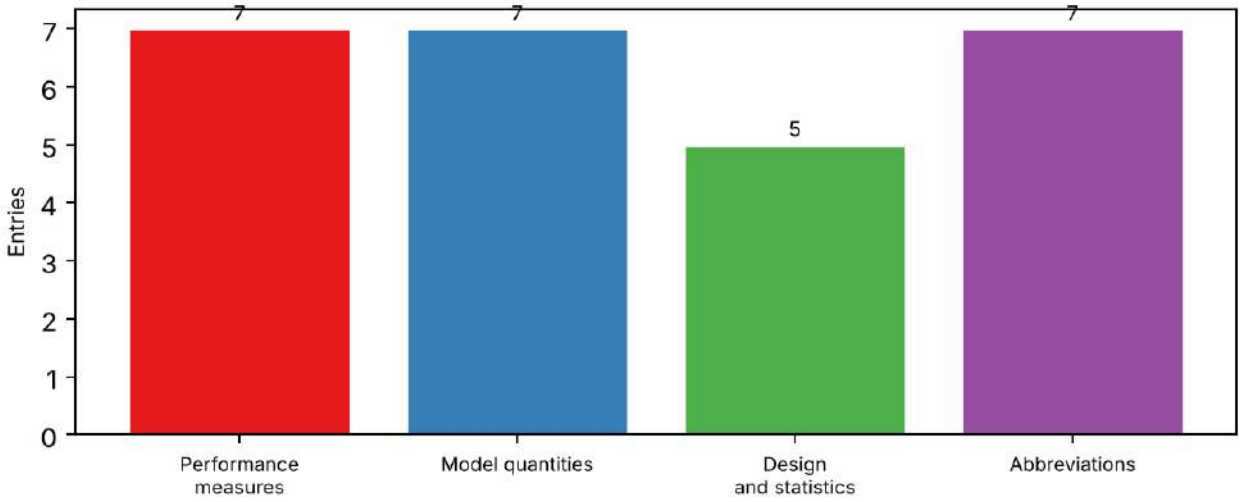

Figure 24: Number of nomenclature entries by category

Table 22: Nomenclature. Symbols, variables and abbreviations used in this paper, with their definition, admissible range and place of introduction

Symbol	Definition	Range or units	Introduced in
M(·)	Task metric (macro-F1 for T1 and T3, AUPRC for T2)	[0, 1]	Eq. (1)
CG	Complementarity gain over the best unimodal model	[-1, 1]	Eq. (1)
MMR	Missing-modality retention	[0, ...)	Eq. (2)
CDD	Cross-dataset drop	[-1, 1]	Eq. (3)
ECE	Expected calibration error	[0, 1]	Eq. (4)
FG	Fairness gap across subgroups	[0, 1]	Eq. (5)
MRI	Modality reliance index	[0, 1]	Eq. (10)
z_h, z_b	Heart and brain embeddings	R^d	Eq. (6)
τ	Temperature of the contrastive loss	0.07	Eq. (6)
L_{NCE}	Contrastive (InfoNCE) loss	R^+	Eq. (6)
Q, K, V	Query, key and value matrices of cross-attention	$R^{n\times d}$	Eq. (7)
m_h, m_b	Presence masks of the ECG and EEG inputs	{0, 1}	Eq. (8)
p	Modality-dropout probability	0.30	Eq. (8)
λ	Weight of the contrastive term	0.5	Eq. (8)
d_s	Per-subject paired difference in the metric	[-1, 1]	Eq. (9)
δ, σ_d	Target difference and SD of paired differences	0.02; 0.06	Eq. (9)
n	Number of test subjects	positive integer	Eq. (9)
α	Significance level, Bonferroni over 6 contrasts	0.05 / 6	Eq. (9)
IG	Integrated gradients attribution	R	Eq. (10)
ECG, EEG	Electrocardiogram; electroencephalogram	-	Section 1
HRV	Heart rate variability	-	Table 8
PSG	Polysomnography	-	Section 3.1
GBM	Gradient-boosted trees	-	Table 4
SNR	Signal-to-noise ratio	dB	Table 12
AUPRC	Area under the precision-recall curve	[0, 1]	Table 6
TRIPOD+AI	Transparent reporting of prediction models, AI extension	-	Section 4.1

VII. CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH

A research program for the evaluation of the added value of multimodal ECG–EEG deep learning has been proposed in this paper. The terms of complementarity, robustness, transfer, calibration, fairness and reliance were operationalized as metrics, the models and hyperparameter search budget, splitting, stress testing, thresholds, and number of samples needed to achieve a minimal detectable effect were all pre-specified for each task, along with the meaning of each result that could occur.

A. Limitations

Cohort. The main cohort is a community sample of older individuals, some of whom were referred for sleep disordered breathing or recruited into a population study. Findings do not generalize to young subjects or other clinical cohorts.

Labels. Sleep stages and arousals are rated by humans, and inter-scorer agreement is imperfect. Affect labels are self-reported. Model performance is constrained by labeling quality.

Montages and devices. The signals are from clinical polysomnography and, in the case of affect, from low-cost wearable devices. Findings may not generalize to other montages, including the dry electrode and single lead devices used for consumers.

Small affect datasets. The affect task has a limited number of subjects, and findings with p-values above the minimum detectable difference are exploratory.

Planning values. Sizes of the datasets, signal properties and the variance of paired differences are planning values, which are confirmed at data access, and the sample size arguments depend on these values.

Not a clinical assessment. The project analyzes signal fusion on retrospective data. It does not address clinical relevance, workflow or outcomes.

B. Future Research

There are three possible future research directions. Firstly, prospective and field validation of the algorithms in the clinical setting where one of the signals is expected to be missing, like wearable ECG with a limited-channel EEG headband. Secondly, clinical applications, for instance whether overnight heart-brain features are associated with future cardiovascular and cognitive events, which would need outcome linked cohorts and a different protocol. Thirdly, extension of the bridge algorithm to other physiological signal pairs, and investigation of whether contrastive alignment of representations improves robustness in the setting of limited labelled subjects.

DECLARATIONS

Conflicts of interest

The authors declare that they have no conflicts of interest.

Data and Code availability

Code, results, configurations and hashed split manifest are open. Other artifacts are released at the levels shown in Table 16.

Declaration of Generative AI and AI-Assisted Technologies

The authors declare that generative AI and AI-assisted technologies were used only for language improvement, proofreading, and editing, where applicable. The authors take full responsibility for the accuracy, originality, and integrity of the content of this manuscript.

Ethics

The study was conducted in accordance with applicable ethical guidelines

REFERENCES

[1] M. A. Samuels, "The brain–heart connection," *Circulation*, vol. 116, no. 1, pp. 77–84, 2007. Available from: https://doi.org/10.1161/CIRCULATIONAHA.106.678995

[2] Silvani, G. Calandra-Buonaura, R. A. L. Dampney, and P. Cortelli, "Brain–heart interactions: Physiology and clinical implications," *Philos. Trans. R. Soc. A*, vol. 374, no. 2067, Art. no. 20150181, 2016. Available from: https://doi.org/10.1098/rsta.2015.0181

[3] Z. Chen, P. Venkat, D. Seyfried, M. Chopp, T. Yan, and J. Chen, "Brain–heart interaction: Cardiac complications after stroke," *Circ. Res.*, vol. 121, no. 4, pp. 451–468, 2017. Available from: https://doi.org/10.1161/CIRCRESAHA.117.311170

[4] Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, pp. 65–69, 2019. Available from: https://www.nature.com/articles/s41591-018-0268-3

[5] H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nat. Commun.*, vol. 11, Art. no. 1760, 2020. Available from: https://www.nature.com/articles/s41467-020-15432-4

[6] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017. Available from: https://doi.org/10.1002/hbm.23730

[7] V. J. Lawhern et al., "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Art. no. 056013, 2018. Available from: https://iopscience.iop.org/article/10.1088/1741-2552/aace8c/meta

[8] Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, 2017. Available from: https://ieeexplore.ieee.org/abstract/document/7961240

[9] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, 2019. Available from: https://ieeexplore.ieee.org/abstract/document/8631195

[10] N. Sridhar et al., "Deep learning for automated sleep staging using instantaneous heart rate," *npj Digit. Med.*, vol. 3, Art. no. 106, 2020. Available from: https://www.nature.com/articles/s41746-020-0291-x

[11] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, 2018. Available from: https://ieeexplore.ieee.org/abstract/document/7887697

[12] A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood

research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, 2021. Available from: https://ieeexplore.ieee.org/abstract/document/8554112

[13] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1692–1706, 2016. Available from: https://ieeexplore.ieee.org/abstract/document/7169562

[14] Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2302–2310, 2021. Available from: https://doi.org/10.1609/aaai.v35i3.16330

[15] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *GigaScience*, vol. 6, no. 5, pp. 1–9, 2017. Available from: https://doi.org/10.1093/gigascience/gix019

[16] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023. Available from: https://www.cell.com/patterns/fulltext/S2666-3899(23)00159-9

[17] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996. Available from: https://doi.org/10.1161/01.CIR.93.5.1043

[18] Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, Art. no. 031001, 2019. Available from: https://doi.org/10.1088/1741-2552/ab0ab5

[19] Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-Time: A fully convolutional network for time series segmentation applied to sleep staging," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019. Available from: https://proceedings.neurips.cc/paper_files/paper/2019/hash/57bafb2c2dfeefba931bb03a835b1fa9-Abstract.html

[20] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019. Available from: https://ieeexplore.ieee.org/abstract/document/8269806

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778. Available from: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794. Available from: https://doi.org/10.1145/2939672.2939785

[23] R. B. Berry et al., *The AASM Manual for the Scoring of Sleep and Associated Events*, Version 2.6. Darien, IL, USA: American Academy of Sleep Medicine, 2020. Available from: https://cir.nii.ac.jp/crid/1370017279880192913

[24] Bonnet et al., "EEG arousals: Scoring rules and examples," *Sleep*, vol. 15, no. 2, pp. 173–184, 1992. Available from: https://doi.org/10.1093/sleep/15.2.174

[25] S. F. Quan et al., "The Sleep Heart Health Study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997. Available from: https://doi.org/10.1093/sleep/20.12.1077

[26] G.-Q. Zhang et al., "The National Sleep Research Resource: Towards a sleep data commons," *J. Amer. Med. Informat. Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018. Available from: https://doi.org/10.1093/jamia/ocy064

[27] M. M. Ghassemi et al., "You snooze, you win: The PhysioNet/Computing in Cardiology Challenge 2018," in *Proc. Comput. Cardiol.*, vol. 45, 2018. Available from: https://ieeexplore.ieee.org/abstract/document/8743916

[28] L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000. Available from: https://cir.nii.ac.jp/crid/1573387449899608448

[29] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696. Available from: https://shorturl.at/qJoa5

[30] Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[31] van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv:1807.03748, 2018. Available from: https://doi.org/10.48550/arXiv.1807.03748

[32] Kiyasseh, T. Zhu, and D. A. Clifton, "CLOCS: Contrastive learning of cardiac signals across space, time, and patients," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, pp. 5606–5615, 2021. Available from: https://proceedings.mlr.press/v139/kiyasseh21a.html

[33] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 70, pp. 1321–1330, 2017. Available from: https://proceedings.mlr.press/v70/guo17a.html

[34] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nat. Med.*, vol. 27, no. 12, pp. 2176–2182, 2021. Available from: https://www.nature.com/articles/s41591-021-01595-0

[35] S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Art. no. e078378, 2024. Available from: https://doi.org/10.1136/bmj-2023-078378

[36] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015. Available from: https://doi.org/10.1371/journal.pone.0118432

[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960. Available from: https://doi.org/10.1371/journal.pone.0118432

[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006. Available from: https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X

[39] W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950. Available from: https://shorturl.at/c85HT

[40] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, 1985. Available from: https://ieeexplore.ieee.org/abstract/document/4122029

[41] Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019. Available from: https://doi.org/10.48550/arXiv.1711.05101

[42] Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum, 1988. Available from: https://shorturl.at/JPQVk

[43] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979. Available from: https://www.jstor.org/stable/4615733

[44] Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945. Available from: https://doi.org/10.2307/3001968

[45] Efron, "Better bootstrap confidence intervals," *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 171–185, 1987. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478410

[46] Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006. Available from: https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[47] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 70, pp. 3319–3328, 2017. Available from: http://proceedings.mlr.press/v70/sundararajan17a.html

[48] Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), *Off. J. Eur. Union*, L 119, 2016. Available from: https://www.disit.org/bellini-sekm/slides-sekm-aa19-20-10.pdf