Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Uncertainty-Aware Neuroprognostication Pipeline Using Longitudinal Brain Continuity Index and Leave-One-Center-Out Validation

Selay Koyuncu

selay.koyuncu@tedu.edu.tr

TED University https://orcid.org/0009-0000-8005-759X

Pardis Azemikhah

Atılım University https://orcid.org/0009-0003-7271-5961

Research Article

Keywords: Brain Continuity Index, EEG, neuroprognostication, multicenter validation, uncertainty estimation, Monte Carlo dropout, selective prediction

Posted Date: March 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9096286/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

An Uncertainty-Aware Neuroprognostication Pipeline Using Longitudinal Brain Continuity Index and Leave-One-Center-Out Validation

Selay Koyuncu[1*†] and Pardis Azemikhah[2†]

[1*]Applied Data Science (M.Sc.), TED University, Ziya Gökalp Caddesi No:48, Kolej, Ankara, 06420, Çankaya, Turkey.
[2]Computer Engineering (M.Sc.), Atılım University, Kızılcaşar Mahallesi, 1184. Cadde No:13, İncek, Ankara, 06830, Gölbaşı, Turkey.

*Corresponding author(s). E-mail(s): selay.koyuncu@tedu.edu.tr;
Contributing authors: azemikhah.pardis@student.atilim.edu.tr;
†These authors contributed equally to this work.

Abstract

Neuroprognostication in intensive care is high-stakes, where overconfident errors may contribute to premature withdrawal of life-sustaining therapy. To improve reproducibility and efficiency over raw EEG modeling, we propose an uncertainty-aware multimodal framework based on the Brain Continuity Index (BCI), a compact EEG-derived representation of background continuity. A dual-branch late-fusion CNN–BiGRU integrates 72-hour longitudinal BCI trajectories with static clinical covariates. In a hospital-level leave-one-center-out evaluation on 607 patients, the model achieved a pooled AUROC of 0.76 and F1-score of 0.80. We observe a performance ceiling associated with the compressed BCI representation: deep sequence modeling was statistically equivalent to classical ensembles ($\boldsymbol{p = 0.413}$), suggesting most prognostic signal is captured by summary-level dynamics. Using Monte Carlo dropout, predictive entropy enabled selective prediction, reaching ≥85% accuracy on high-confidence cases while deferring uncertain cases for expert review. These results highlight multicenter robustness and uncertainty-aware triage as key elements for trustworthy AI in neurocritical care.

Keywords: Brain Continuity Index; EEG; neuroprognostication; multicenter validation; uncertainty estimation; Monte Carlo dropout; selective prediction

Author biographies

Selay Koyuncu is an M.Sc. student in Applied Data Science at TED University (Ankara, Turkey). Pardis Azemikhah is an M.Sc. student in Computer Engineering at Atılım University (Ankara, Turkey).

1 Introduction

Clinical decision-making in the intensive care unit (ICU) is time-critical and uncertainty-driven; early choices can have irreversible consequences [1–3]. Post–cardiac arrest neuroprognostication is a prototypical example: many patients achieve return of spontaneous circulation (ROSC) yet remain comatose, requiring outcome estimates that inform care goals and communication [4]. Critically, false pessimistic predictions may trigger premature withdrawal of life-sustaining therapy (WLST), so clinically useful tools must minimize false positives and communicate reliability [5, 6].

Electroencephalography (EEG) is central because it is bedside-accessible, continuous, and reflects post-anoxic brain dynamics. Established prognostic markers include background continuity, suppression/burst-suppression, and reactivity [7–10]. Quantitative EEG (qEEG) features (e.g., spectral and connectivity measures) further support scalable, multicenter outcome prediction [11]. The International Cardiac Arrest Research (I-CARE) consortium enabled rigorous benchmarking through a large multicenter EEG and clinical database [12]. The George B. Moody PhysioNet Challenge 2023 emphasized safety by evaluating at a conservative operating point (TPR at fixed FPR=0.05), discouraging overconfident pessimistic predictions [13].

Direct modeling of raw multichannel EEG remains computationally expensive and difficult to reproduce. To reduce this burden, we use the Brain Continuity Index (BCI) released via Dryad [14]: a compact longitudinal feature matrix that summarizes background continuity at 10-second resolution for 607 I-CARE subjects. BCI is a validated qEEG biomarker defined as the fraction of time a segment remains above a 10μV suppression threshold for more than 0.5 seconds [14]. Using this 52 MB representation avoids processing the 1.5–3.9 TB raw waveform corpus while retaining clinically interpretable trajectory dynamics over 72 hours (25,920 steps per patient).

Despite progress, key challenges persist: EEG interpretation requires expertise and remains variable; post–cardiac arrest data are often imbalanced; cross-hospital generalization is fragile under institutional shift; and uncertainty quantification and fairness auditing are still under-integrated in neuroprognostic pipelines [2, 4, 6]. Classical models typically rely on engineered summaries (e.g., means, variances, suppression fractions, slopes) coupled with logistic regression or tree ensembles, offering efficiency and interpretability but limited temporal expressiveness. Deep models (CNNs, recurrent networks) can learn temporal representations directly, yet may overfit modest clinical cohorts and often lack strict external validation.

Here, we study multicenter neuroprognostication using publicly available BCI trajectories [14]. We propose a multimodal CNN–BiGRU that fuses 72-hour BCI sequences with six clinical covariates (age, sex, ROSC duration, OHCA, initial rhythm, TTM). To ensure external validity, we use Leave-One-Hospital-Out (LOHO) evaluation, with threshold selection on validation data within training hospitals. We

benchmark against classical baselines (XGBoost, random forest, logistic regression) trained on 27 engineered features (21 temporal summaries + 6 covariates) under identical splits.

Beyond discrimination, we report calibration (ECE, Brier), subgroup fairness (sex and age strata), and uncertainty-aware selective prediction via Monte Carlo dropout. This design allows us to test whether end-to-end sequence modeling provides meaningful gains over engineered summaries under a compressed EEG abstraction, while producing deployment-relevant evidence for trustworthy AI in neurocritical care.

Paper outline and positioning.

High-performing I-CARE systems often rely on heavy raw-EEG preprocessing and rich multimodal features, optimized for extremely low false-positive regimes [13]. In contrast, we examine what can be achieved from a compact, public abstraction (BCI), emphasizing strict hospital isolation, calibration, uncertainty-aware triage, and demographic auditing. The next section reviews related work and motivates our reproducible BCI-focused framework.

2 Background and Related Work

2.1 I-CARE and the PhysioNet/George B. Moody Challenge

Because false pessimistic predictions can contribute to premature withdrawal of life-sustaining therapy, the I-CARE Challenge prioritized safety by evaluating models at a conservative operating point: true positive rate at a fixed false positive rate of 0.05. This setup encourages robust performance under uncertainty rather than optimizing average accuracy. High-scoring submissions (e.g., deep transfer learning pipelines such as ComaToss) demonstrate that competitive performance is achievable under stringent low-FPR constraints when operating on rich raw-EEG representations and strong augmentation strategies [13, 15].

2.2 Feature-based machine learning on EEG/qEEG

A common and effective approach is to extract quantitative EEG features (e.g., PSD-band power, coherence/connectivity, burst-suppression metrics, and nonlinear complexity measures) and train classical learners such as gradient boosting, random forests, or SVMs. Across I-CARE-related studies, engineered features combined with boosted trees frequently yield strong discrimination, and connectivity/complexity features can be especially influential [11, 16–19]. Multimodal fusion with clinical covariates (and ECG/HRV when available) is repeatedly reported to increase reliability and improve conservative operating points [12, 13, 20–22].

2.3 Deep learning for temporal EEG dynamics

Deep architectures aim to learn hierarchical temporal representations and long-range dependencies more directly than handcrafted summaries. CNN-based encoders are effective for local motifs, while recurrent and transformer-based models capture longer

temporal structure; recent I-CARE lines of work include multiscale CNN–(Bi)LSTM pipelines, spectrogram-based CNNs, and transformer variants [23–26]. A recurring observation is that temporal evolution matters: performance tends to improve at later time windows (e.g., 24–48h or beyond), suggesting that time-aware modeling can better reflect stabilization and evolving physiology [23].

2.4 Trustworthy AI gaps: calibration, uncertainty, and transparency

In high-stakes neuroprognostication, discrimination alone is insufficient. The literature increasingly emphasizes probabilistic reliability (calibration), robustness to inter-center shift, explicit uncertainty representation, and transparency/interpretability to support clinician trust and responsible deployment [2, 4, 6, 27]. Explainability tools (e.g., SHAP-type analyses) and calibration reporting are frequently recommended as complements to raw performance metrics, particularly under WLST-related ethical constraints [27, 28].

2.5 Our positioning: reproducible modeling under a compressed EEG representation (BCI)

Most high-performing I-CARE pipelines operate on raw multichannel EEG (often multimodal) and depend on heavy preprocessing and large-scale feature extraction. In contrast, the publicly released Brain Continuity Index (BCI) provides a compact, standardized abstraction of background continuity dynamics at 10-second resolution [14]. This enables reproducible, resource-efficient benchmarking of (i) sequence modeling versus engineered summaries under strict hospital isolation, and (ii) deployment-relevant properties such as calibration, selective prediction via uncertainty, and subgroup auditing. Accordingly, our goal is not leaderboard optimization but transparent assessment of what can (and cannot) be achieved from the BCI abstraction under multicenter validation.

A critical implication of adopting a compressed EEG abstraction such as BCI is the potential existence of a representational performance ceiling. Because BCI encodes background continuity as a single-channel scalar trajectory rather than preserving full multichannel spectral and spatial structure, its informational bandwidth is inherently constrained. Under such compression, carefully engineered temporal descriptors (e.g., variance, suppression burden, and rolling-window volatility) may already capture a large proportion of the prognostic signal, limiting the marginal benefit of deeper hierarchical sequence modeling. Accordingly, in this study we explicitly test whether end-to-end representation learning from the full 25,920-step BCI trajectory provides statistically meaningful gains beyond engineered summaries, rather than assuming architectural superiority a priori.

3 Methods

3.1 Dataset and Cohort Description

We utilized the publicly released Brain Continuity Index (BCI) dataset made available through the Dryad repository by the I-CARE consortium [14]. The dataset comprises a multi-center cohort of 607 adult patients who sustained anoxic brain injury following cardiac arrest. Patients were recruited across five geographically distinct hospitals (A, B, D, E, F), thereby enabling institutional-level generalization analysis. For each patient, a longitudinal BCI trajectory spanning the first 72 hours of intensive care was available. BCI values were sampled at 10-second resolution, yielding a fixed-length time series of 25,920 sequential measurements per subject. The BCI is a scalar measure bounded between 0 and 1, where lower values reflect EEG background suppression and higher values indicate restoration of continuous cortical activity. By modeling these trajectories, we aim to capture both the degree and temporal evolution of neurological recovery. In addition to the temporal signal, six static clinical covariates were included: age (normalized), sex, duration of return of spontaneous circulation (ROSC), out-of-hospital cardiac arrest status (OHCA), initial shockable rhythm, and application of targeted temperature management (TTM). The clinical endpoint was neurological outcome assessed using the Cerebral Performance Category (CPC) scale. Outcomes were dichotomized into good (CPC 1-2) and poor (CPC $\geq$3), defining a binary prediction task. Importantly, no raw EEG waveforms were accessed or processed in this study. All modeling was conducted exclusively using the standardized BCI trajectories provided in the public release, thereby ensuring reproducibility and minimizing computational barriers to replication. All preprocessing, model training, evaluation, and statistical analyses were implemented in Python (PyTorch, scikit-learn, and XGBoost). To ensure full computational reproducibility, the complete experimental pipeline, configuration files, and trained model artifacts are publicly available in our GitHub repository [29]. The repository contains scripts for Leave-One-Hospital-Out (LOHO) splitting, feature engineering, calibration analysis, uncertainty estimation via Monte Carlo dropout, and subgroup fairness evaluation, enabling independent verification of all reported results.

3.2 Cross-Hospital Validation Strategy

To rigorously evaluate external generalization and prevent institutional data leakage, we adopted a Leave-One-Hospital-Out (LOHO) cross-validation framework. In each of five folds, one entire hospital was held out as a fully unseen test environment. The remaining four hospitals were pooled, and their subjects were randomly partitioned at the patient level into an 80% training subset and a 20% validation subset. This strategy preserves strict geographic isolation for test evaluation while allowing robust hyperparameter selection and threshold tuning within the pooled training hospitals. Crucially, the held-out hospital was never exposed during model optimization, early stopping, temperature scaling, or threshold calibration. No patient appeared in more than one partition within any fold. Performance was reported using two complementary approaches. First, fold-wise metrics were computed independently on each test hospital and averaged (macro mean-of-fold). Second, predictions from all five test

folds were concatenated to produce pooled metrics across the full cohort. The former reflects hospital-level stability, while the latter quantifies overall discrimination capacity across institutions.

3.3 Deep Multimodal CNN–BiGRU Architecture

To directly model longitudinal BCI dynamics, we designed a multimodal deep learning architecture combining temporal representation learning with structured clinical information. The temporal branch receives the full 25,920-step BCI sequence as input. This one-dimensional signal is processed through three stacked 1D convolutional layers with kernel sizes of 10, 8, and 5 and corresponding strides of 5, 4, and 5. Each convolutional layer is followed by batch normalization, rectified linear unit activation, and dropout ($p = 0.2$) to promote generalization. These convolutional filters extract local temporal motifs, including suppression bursts and continuity transitions. The convolutional output is then passed to a bidirectional gated recurrent unit (BiGRU) layer with hidden dimension 64. The BiGRU captures long-range temporal dependencies and recovery trajectories in both forward and backward directions, enabling the model to integrate early suppression and later continuity patterns. Adaptive global average pooling compresses the recurrent output into a fixed 64-dimensional temporal embedding. In parallel, the six static clinical covariates are processed through a two-layer multilayer perceptron (MLP) with hidden dimension 32 and dropout regularization. This branch produces a 32-dimensional demographic embedding. The temporal and static embeddings are concatenated and passed through a fusion MLP ($96 \rightarrow 32 \rightarrow 1$), producing a scalar logit. A sigmoid activation transforms this logit into a probability estimate of poor neurological outcome. To mitigate class imbalance, we trained the network using Focal Loss with focusing parameter $\gamma = 2.0$ and class weighting parameter $\alpha = 0.25$. Focal Loss down-weights well-classified examples and emphasizes difficult cases, thereby reducing majority-class dominance during gradient updates. Rather than adopting a fixed 0.5 classification threshold, we determined the operational decision threshold using the Youden Index computed on the validation subset of each fold. The selected threshold was then applied to the held-out test hospital. The summarized architecture of the hybrid model can be seen in Figure 1.

3.4 Classical Machine Learning Baselines and Feature Engineering

To contextualize the performance of deep representation learning, we constructed three classical machine learning baselines: XGBoost, Random Forest, and L2-regularized Logistic Regression. Because these models cannot efficiently process long temporal sequences directly, we engineered a compact set of 27 features per patient. Twenty-one features were derived from the 72-hour BCI trajectory. These included global summary statistics (mean, minimum, maximum, standard deviation, and global linear slope), suppression burden metrics (fraction of time below amplitude thresholds of 0.1, 0.2, and 0.3), and time-to-continuity. Additionally, to capture burst-suppression volatility across multiple temporal scales, rolling window statistics were computed at 1-minute, 5-minute, and 30-minute resolutions. For each window scale, maximum and

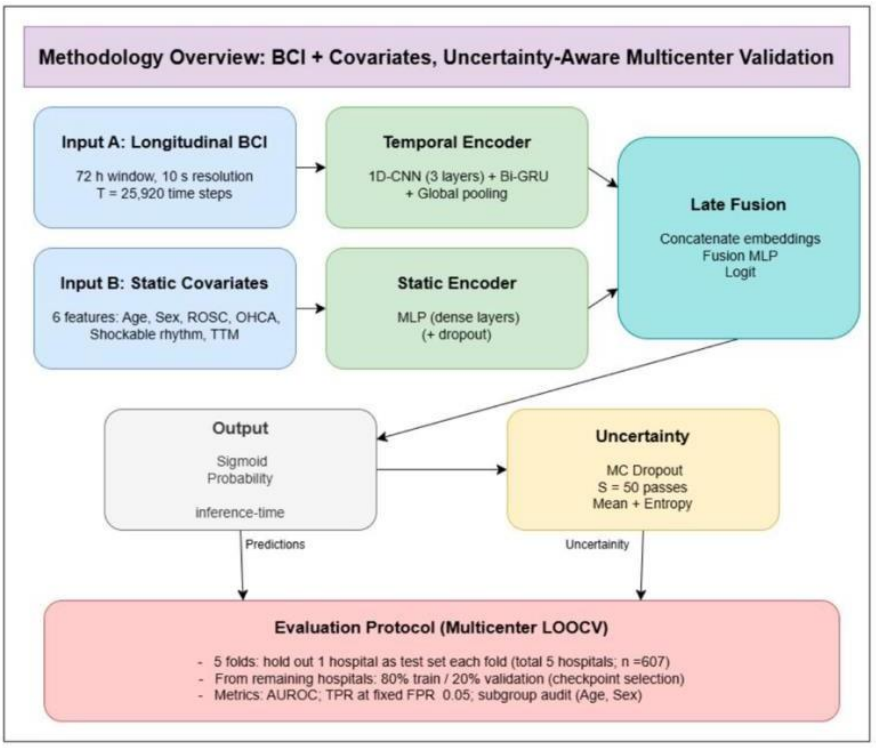


Fig. 1 Methodology overview: longitudinal BCI and static covariates are encoded by a CNN–BiGRU temporal encoder and an MLP static encoder, fused via late fusion; uncertainty is estimated using MC dropout; evaluation uses leave-one-hospital-out validation.

minimum rolling means and variances were extracted. Notably, the 30-minute rolling variance maximum emerged as one of the most predictive features under gain-based importance analysis. These 21 temporal features were concatenated with the same six static clinical covariates used in the deep model, yielding a 27-dimensional feature vector for classical training. All baseline models were trained and validated under identical LOHO splits to ensure fair comparison.

3.5 Calibration Analysis

Beyond discrimination metrics, we evaluated probabilistic calibration using Expected Calibration Error (ECE) with 15 equal-width bins, Brier score, and negative log-likelihood. ECE measures the absolute difference between predicted confidence and empirical accuracy within each probability bin. Temperature scaling was optionally applied as a post-hoc calibration technique. The temperature parameter was optimized exclusively on validation logits within each fold and subsequently applied to test logits, preserving isolation of the held-out hospital.

3.6 Uncertainty Estimation and Selective Prediction

To quantify predictive uncertainty, we employed Monte Carlo dropout at inference time with 50 stochastic forward passes. The predictive mean was computed across stochastic samples, and predictive entropy was used as an uncertainty measure. Selective prediction performance was evaluated using risk-coverage curves. Predictions were sorted by uncertainty, and classification accuracy was computed as progressively uncertain samples were rejected. We report baseline accuracy at full coverage, maximum achievable accuracy under abstention, and coverage thresholds required to reach high-confidence performance levels (e.g., $\geq$0.85 accuracy).

3.7 Fairness Evaluation

To assess demographic fairness, performance was stratified by sex (female vs male) and age (above vs below median), as well as by intersectional combinations of these categories. For each subgroup, AUROC, PR-AUC, F1-score, false positive rate, false negative rate, and calibration metrics were computed. Bootstrap resampling (n = 1000) was used to generate 95% confidence intervals for subgroup AUROC and F1 values. Disparity metrics were defined as the difference between subgroup performances, and statistical significance was inferred when confidence intervals excluded zero.

3.8 Statistical Testing and Challenge Metric

To compare model discrimination, we used the DeLong test for correlated AUROC curves on pooled out-of-bag predictions. Operational performance differences in F1-score were evaluated using paired bootstrap resampling. For benchmarking against prior neuroprognostication studies, we additionally computed the challenge score defined as the true positive rate achieved at a fixed false positive rate ≤ 0.05. This metric reflects clinically conservative operating regimes emphasizing minimization of false pessimistic predictions.

4 Results

4.1 Discrimination and Clinical Operating Performance

Across the five leave-one-hospital-out folds, the Multimodal CNN-BiGRU demonstrated consistent generalization performance when evaluated on completely unseen institutions. Operating strictly on held-out hospital test sets, the model achieved a pooled AUROC of 0.760 (95% CI: 0.72–0.79) and a pooled F1-score of 0.797 under validation-derived operational thresholding. The mean-of-fold AUROC was 0.742 ± 0.070, reflecting moderate variability across hospitals A, B, D, E, and F, but no catastrophic degradation in any individual institution. These results indicate stable multicenter discrimination under institutional domain shift. Importantly, hospital-level isolation ensured that no data from the held-out institution influenced feature extraction, threshold optimization, or calibration procedures. Strikingly, the classical XGBoost baseline, trained exclusively on the 27 engineered features, achieved a

Table 1 Performance comparison of all models under hospital-level leave-one-hospital-out cross-validation. Mean metrics represent the average across five held-out hospitals; pooled metrics use all out-of-bag predictions combined across folds. Confidence intervals are 95% bootstrap intervals (1,000 resamples).

Model	Mean AUROC	Pooled AUROC	Pooled AUROC 95% CI	Pooled AUPRC	Mean F1	Pooled F1	Pooled F1 95% CI	Pooled ECE	Pooled Brier
Multimodal CNN–BiGRU	0.74	0.76	0.721–0.795	0.82	0.80	0.80	0.766–0.825	0.15	0.20
XGBoost	0.69	0.72	0.682–0.766	0.81	0.82	0.81	0.783–0.835	0.12	0.21
Random Forest	0.68	0.70	0.646–0.736	0.78	0.79	0.77	0.740–0.798	0.11	0.22
Logistic Regression	0.71	0.72	0.676–0.758	0.78	0.77	0.75	0.720–0.785	0.11	0.21

pooled AUROC of 0.724 (95% CI: 0.68-0.76) and a pooled F1-score of 0.810, numerically surpassing the deep model's F1-score. Random Forest and L2-regularized Logistic Regression performed modestly worse, with lower pooled AUROC values and inferior F1 scores. To determine whether direct modeling of the 25,920-step BCI trajectory provided statistically meaningful benefit over engineered temporal summaries, we performed formal statistical testing. The DeLong test comparing pooled AUROC between the Multimodal CNN-BiGRU and XGBoost yielded $p = 0.413$, indicating no statistically significant difference in discrimination. Likewise, a paired bootstrap test (1,000 resamples) comparing F1-scores produced $p = 0.378$, confirming statistical equivalence in operational performance. Taken together, these results suggest that representation learning over the full high-resolution BCI sequence does not confer a statistically significant advantage over well-designed temporal summary statistics in this dataset. Table 1 breaks down the performance of each model. The pooled ROC curves comparing the Multimodal CNN–BiGRU and XGBoost are shown in Figure 2, with fold-wise hospital-level ROC analyses included in Supplementary Figure S1 to demonstrate performance consistency under strict institutional isolation.

4.2 Feature Efficacy and Interpretability

To interrogate the predictive structure of the classical model, we examined feature importance rankings from XGBoost using average gain across all LOHO folds. The most influential predictors consistently included:

- Age
- 30-minute rolling variance maximum
- Global BCI standard deviation
- ROSC duration

The dominance of rolling variance metrics suggests that high-amplitude volatility at longer time scales is strongly associated with outcome class separation. This finding aligns with clinical observations that burst-suppression instability and prolonged suppression variability are prognostically informative. Notably, the deep CNN-BiGRU likely learns analogous representations internally, as convolutional filters and recurrent units are well-suited to capture variance and temporal slope patterns. The statistical parity between deep and classical approaches reinforces the hypothesis that a limited number of summary statistics encapsulate the majority of prognostic signal in the BCI representation. Figure 3 presents the top 15 XGBoost feature importances averaged across leave-one-hospital-out folds.

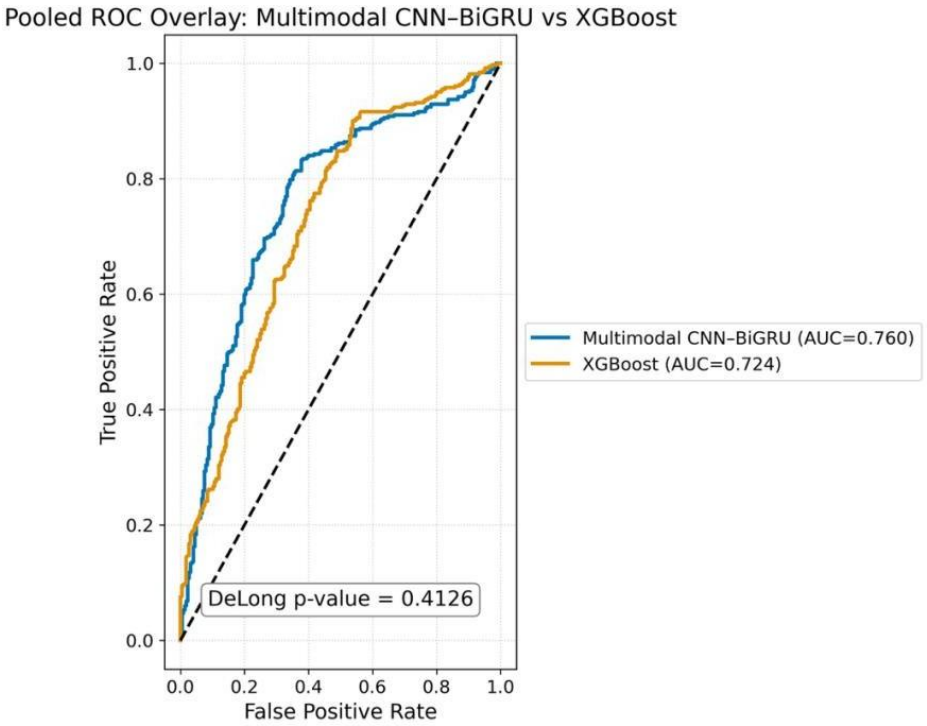


Fig. 2 Pooled ROC overlay comparing Multimodal CNN–BiGRU and XGBoost under strict hospital isolation (DeLong p-value shown).

4.3 Fairness and Subgroup Disparities

To evaluate demographic fairness under multicenter deployment, we stratified performance by sex (female vs male) and age (above vs below median), including intersectional splits. The Multimodal CNN-BiGRU demonstrated minimal discrimination disparity across sex groups. The bootstrapped 95% confidence interval for ΔAUROC (Female - Male) crossed zero, indicating no statistically significant difference in discrimination between male and female patients. Age-based splits exhibited slightly larger variance in AUROC and F1-score, with older patients demonstrating modestly lower AUROC but higher F1-scores due to outcome prevalence distribution differences. However, bootstrapped 95% confidence intervals for ΔAUROC between age strata also crossed zero, preventing declaration of systematic age-based bias. Intersectional subgroup evaluation similarly revealed overlapping confidence intervals. While point estimates varied slightly across subgroups, no statistically robust disparity was identified. These findings suggest that under strict hospital-level validation, the model does not exhibit measurable demographic discrimination in terms of AUROC or F1-score. Table 2 presents the performance of each subgroup with numbers and figure 4 showcases the error bars for each.

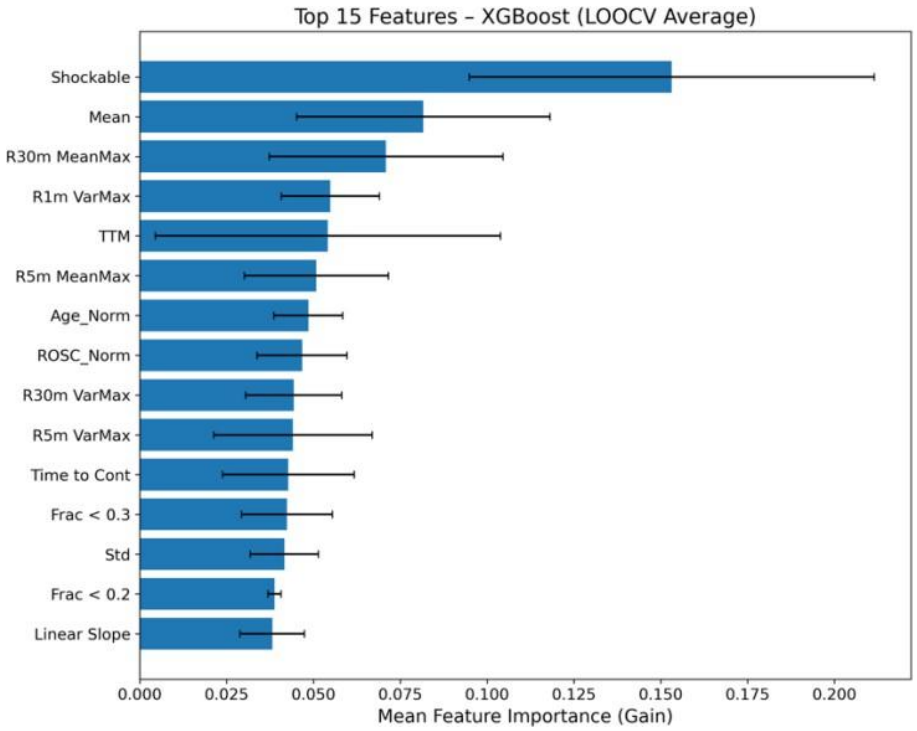


Fig. 3 Top 15 XGBoost feature importances (average gain across LOHO folds).

4.4 Calibration and Uncertainty-Aware Prediction

4.4.1 Reliability and Calibration

Calibration analysis revealed moderate misalignment between predicted probabilities and empirical risk prior to temperature scaling. The pooled Expected Calibration Error (ECE) for the Multimodal CNN-BiGRU was 0.146, compared to 0.122 for XGBoost. Both models demonstrated slight underconfidence in mid-probability ranges. Post-hoc temperature scaling, performed exclusively on validation data, reduced pooled ECE for the deep model to approximately 0.114, improving probabilistic reliability without altering discrimination. These results indicate that while raw predictive logits are moderately miscalibrated, simple monotonic rescaling effectively restores probabilistic interpretability. As depicted in Figure 5, the pooled reliability diagrams reveal slight probability miscalibration in both the Multimodal CNN-BiGRU and XGBoost prior to post-hoc scaling, particularly within mid-range confidence intervals.

4.4.2 Selective Prediction and Risk-Coverage Trade-Off

Given the high-stakes nature of neurological outcome prediction, we implemented an uncertainty-aware selective prediction framework in which Monte Carlo (MC)

Table 2 Fairness evaluation across sex, age strata, and intersectional subgroups for Multimodal CNN–BiGRU and XGBoost.

Model	Subgroup	N	Prevalence	AUROC	AUROC CI	PRAUC	F1	F1 CI
Multimodal CNN–BiGRU	Female	190	0.67	0.76	0.683–0.838	0.84	0.85	0.806–0.897
Multimodal CNN–BiGRU	Male	417	0.61	0.75	0.706–0.801	0.82	0.77	0.726–0.809
Multimodal CNN–BiGRU	Age_Below_Median	307	0.53	0.77	0.718–0.821	0.79	0.76	0.703–0.802
Multimodal CNN–BiGRU	Age_Above_Median	300	0.73	0.72	0.659–0.792	0.85	0.83	0.785–0.864
Multimodal CNN–BiGRU	Female_Below	101	0.59	0.78	0.686–0.862	0.84	0.80	0.721–0.870
Multimodal CNN–BiGRU	Female_Above	89	0.75	0.72	0.567–0.857	0.85	0.90	0.835–0.943
Multimodal CNN–BiGRU	Male_Below	206	0.50	0.77	0.701–0.833	0.76	0.73	0.660–0.795
Multimodal CNN–BiGRU	Male_Above	211	0.72	0.72	0.642–0.792	0.85	0.80	0.742–0.842
XGBoost	Female	190	0.67	0.70	0.614–0.776	0.82	0.82	0.772–0.869
XGBoost	Male	417	0.61	0.74	0.686–0.790	0.80	0.80	0.766–0.840
XGBoost	Age_Below_Median	307	0.53	0.73	0.668–0.782	0.75	0.75	0.698–0.792
XGBoost	Age_Above_Median	300	0.73	0.67	0.599–0.739	0.84	0.86	0.822–0.891
XGBoost	Female_Below	101	0.59	0.74	0.643–0.830	0.80	0.78	0.694–0.855
XGBoost	Female_Above	89	0.75	0.57	0.436–0.707	0.82	0.86	0.803–0.916
XGBoost	Male_Below	206	0.50	0.73	0.653–0.792	0.72	0.73	0.664–0.791
XGBoost	Male_Above	211	0.72	0.71	0.629–0.780	0.85	0.86	0.816–0.893

Table 3 (continued) Fairness evaluation across sex, age strata, and intersectional subgroups for Multimodal CNN–BiGRU and XGBoost (continued).

Model	Subgroup	Precision	Recall	FPR	FNR	ECE (pre)	Brier
Multimodal CNN–BiGRU	Female	0.84	0.87	0.33	0.13	0.16	0.19
Multimodal CNN–BiGRU	Male	0.75	0.78	0.40	0.22	0.14	0.21
Multimodal CNN–BiGRU	Age_Below_Median	0.74	0.78	0.32	0.22	0.13	0.21
Multimodal CNN–BiGRU	Age_Above_Median	0.82	0.83	0.49	0.17	0.17	0.19
Multimodal CNN–BiGRU	Female_Below	0.79	0.82	0.32	0.18	0.11	0.20
Multimodal CNN–BiGRU	Female_Above	0.88	0.91	0.36	0.09	0.22	0.18
Multimodal CNN–BiGRU	Male_Below	0.71	0.76	0.32	0.24	0.14	0.22
Multimodal CNN–BiGRU	Male_Above	0.79	0.80	0.53	0.20	0.16	0.20
XGBoost	Female	0.77	0.89	0.54	0.11	0.14	0.20
XGBoost	Male	0.71	0.93	0.60	0.07	0.14	0.21
XGBoost	Age_Below_Median	0.67	0.85	0.48	0.15	0.16	0.23
XGBoost	Age_Above_Median	0.77	0.97	0.77	0.03	0.13	0.19
XGBoost	Female_Below	0.76	0.80	0.37	0.20	0.15	0.21
XGBoost	Female_Above	0.77	0.97	0.86	0.03	0.16	0.20
XGBoost	Male_Below	0.63	0.88	0.53	0.13	0.19	0.24
XGBoost	Male_Above	0.77	0.97	0.73	0.03	0.12	0.18

dropout entropy was used as a confidence metric. Rather than forcing a prediction for every patient, the model was allowed to abstain on cases exhibiting high epistemic uncertainty, thereby enabling a principled risk-coverage trade-off.

For each patient x_i, we obtained $S = 50$ stochastic forward passes with dropout active at inference time, yielding predictive probabilities $\hat{p}_s(x_i)$. The final predictive mean was computed as in Eq. 1.

$$\bar{p}(x_i) = \frac{1}{S}\sum_{s=1}^{S} \hat{p}_s(x_i) \tag{1}$$

$$\hat{y}_i = \mathrm{I}[\bar{p}(x_i) \geq \tau] \tag{2}$$

$$H(x_i) = -\bar{p}(x_i)\log \bar{p}(x_i) - (1-\bar{p}(x_i))\log(1-\bar{p}(x_i)) \tag{3}$$

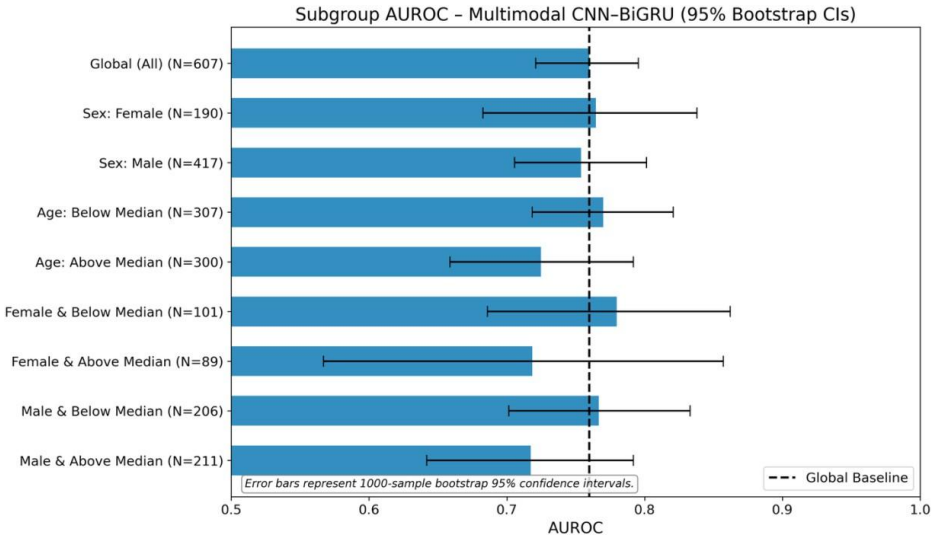


Fig. 4 Subgroup AUROC values for the Multimodal CNN–BiGRU with 95% bootstrap confidence intervals (1,000 resamples).

$$S_\kappa = \{i \mid H(x_i) \leq \kappa\} \tag{4}$$

$$\text{Coverage}(\kappa) = \frac{|S_\kappa|}{N} \tag{5}$$

$$\text{Accuracy}(\kappa) = \frac{1}{|S_\kappa|} \sum_{i \in S_\kappa} \mathbb{I}[\hat{y}_i = y_i] \tag{6}$$

Binary predictions were generated at the operational threshold τ (Eq. 2). Uncertainty was quantified using predictive entropy (Eq. 3). Higher entropy values indicate greater predictive ambiguity. To enable selective prediction, we defined a rejection threshold κ over entropy and retained only cases satisfying Eq. 4. Coverage was defined as the proportion of retained patients (Eq. 5). Accuracy among retained cases was computed as in Eq. 6.

By varying κ from its maximum to minimum value, we obtained the full risk-coverage curve, which characterizes how predictive reliability evolves as low-confidence cases are progressively excluded. At full coverage (Coverage = 1), the Multimodal CNN-BiGRU achieved an overall classification accuracy of 74.0%. As high-entropy cases were increasingly rejected, accuracy improved sharply among the retained subset. The model crossed the clinically meaningful threshold of ≥85% accuracy when coverage was reduced to approximately 0.28, corresponding to abstaining on roughly 72% of the most uncertain predictions. This analysis demonstrates that while full-population deployment yields moderate accuracy, a high-confidence subset of predictions can achieve substantially greater reliability, supporting safe integration within clinician-in-the-loop workflows. This reject-option behavior is particularly relevant for ICU deployment scenarios, where uncertain cases may be escalated for multimodal expert

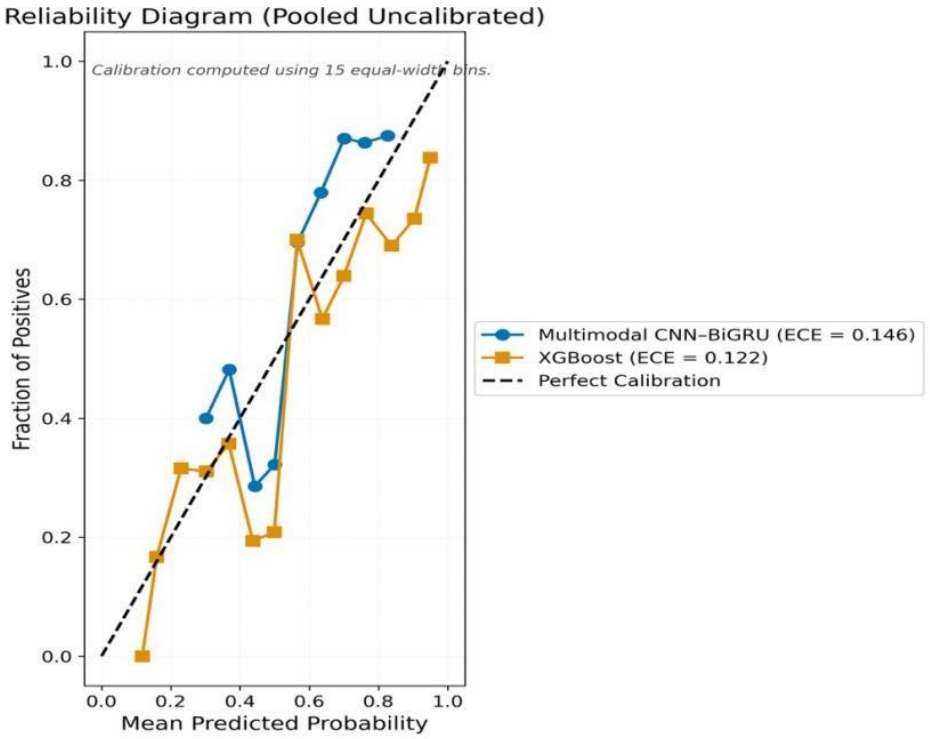


Fig. 5 Pooled reliability diagram comparing Multimodal CNN–BiGRU and XGBoost (uncalibrated; 15 equal-width bins).

review rather than automated classification. Figure 6 illustrates the risk-coverage relationship for the Multimodal CNN-BiGRU, demonstrating how classification accuracy increases as progressively uncertain cases are abstained from under entropy-based rejection.

4.5 Challenge Score Benchmarking and Contextualization

To facilitate benchmarking against prior PhysioNet I-CARE Challenge submissions, we computed the official challenge score defined as the true positive rate (TPR) achieved at a fixed false positive rate (FPR ≤ 0.05). Under strict hospital-level isolation, the Multimodal CNN- BiGRU achieved a challenge score slightly above 0.27. It is important to contextualize this value. High-ranking submissions in the official Challenge leaderboard frequently reported substantially higher scores. However, those solutions operated directly on the full raw I-CARE EEG waveform dataset, often incorporating multi-channel EEG recordings, additional physiological signals, high-dimensional spectral and time-frequency features, extensive handcrafted and hybrid feature sets, explicit optimization targeting the FPR ≤ 0.05 regime. In contrast, the present study deliberately restricted modeling to the publicly released Brain Continuity Index (BCI) trajectories from Dryad. This representation is a compressed, singlechannel longitudinal abstraction of EEG background continuity rather than the

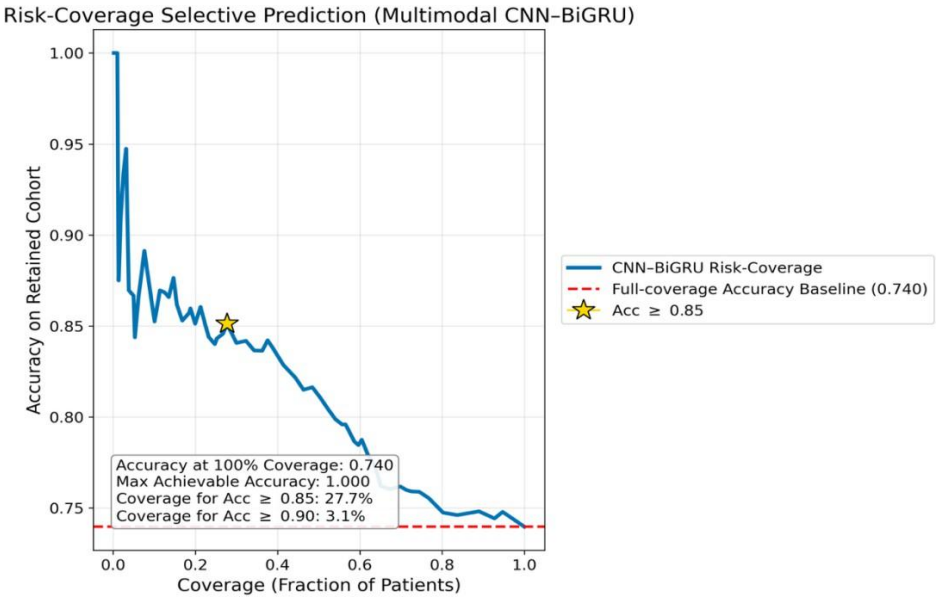


Fig. 6 Risk–coverage curve for uncertainty-based selective prediction using MC dropout entropy (Multimodal CNN–BiGRU).

complete raw multivariate EEG signal. Therefore, the objectives of this work differ fundamentally from leaderboard optimization. Specifically, this study was designed to:

1. Evaluate reproducible modeling using a compact, publicly accessible EEG-derived representation.
2. Compare deep sequence modeling to engineered feature extraction under identical institutional isolation.
3. Quantify calibration, uncertainty, and demographic fairness.
4. Emphasize multicenter robustness over single-metric maximization.

The challenge score is thus reported transparently for benchmarking purposes but should not be interpreted as the primary performance indicator of this framework. Instead, the emphasis of this study lies in multicenter generalizability, statistical equivalence testing, probabilistic calibration, and safe deployment characteristics.

5 Discussion

This multicenter investigation examined whether deep sequence modeling of 72-hour Brain Continuity Index (BCI) trajectories provides meaningful clinical advantage over classical machine learning models trained on carefully engineered summary statistics. Under a strict hospital-level leave-one-center-out evaluation framework with subject-level validation partitioning, the Multimodal CNN-BiGRU achieved a pooled AUROC of 0.760 and an optimized pooled F1-score of 0.797. Importantly, XGBoost trained on 27 engineered features, comprising 21 temporal BCI statistics and 6 static clinical covariates, demonstrated statistical non-inferiority, achieving an AUROC of 0.724

and an F1-score of 0.810. Formal DeLong testing for AUROC and paired bootstrap testing for F1 confirmed no statistically significant performance difference between the deep and classical approaches. An important conceptual finding of this study is the apparent “signal ceiling” inherent to the BCI abstraction. Despite the architectural flexibility of the Multimodal CNN-BiGRU, its discrimination performance was statistically equivalent to that of XGBoost trained on 27 carefully engineered temporal and clinical features (DeLong $p = 0.413$). This suggests that, within the compressed BCI representation, much of the prognostic signal is already captured by relatively simple descriptors such as age, global variance, suppression fractions, and multi-scale rolling volatility metrics. In other words, when the input signal has already been distilled into a one-dimensional continuity index, the marginal benefit of deep hierarchical feature extraction may be limited. The ceiling is not necessarily imposed by model capacity, but by the informational bandwidth of the representation itself. Accordingly, the contribution of the CNN-BiGRU architecture should not be interpreted as achieving higher raw accuracy, but rather as providing a flexible probabilistic modeling framework. Deep sequence models allow end-to-end representation learning, joint embedding with static covariates, and seamless integration of stochastic inference mechanisms such as Monte Carlo dropout. While classical ensemble methods can approximate strong discrimination performance, they do not inherently provide principled epistemic uncertainty estimation within the same unified probabilistic architecture. Thus, under signal compression constraints, the added value of deep learning in this setting lies less in improved scores and more in enhanced modeling extensibility, uncertainty awareness, and deployment readiness. These findings indicate that within the compressed BCI representation, much of the prognostic signal may be captured by relatively simple temporal variance and suppression descriptors. The deep convolutional-recurrent encoder did not substantially exceed the predictive ceiling achieved by rolling window statistics and global volatility measures. This does not imply that deep learning lacks utility; rather, it suggests that the BCI signal abstraction itself may constrain the maximum achievable representational gain. Because the BCI encodes background continuity amplitude rather than raw multichannel EEG waveforms, its informational bandwidth is inherently limited. Under such compression, engineered features approximating variance dynamics, suppression fractions, and trajectory slopes may sufficiently summarize the underlying prognostic structure. Feature importance analysis from XGBoost further supports this interpretation. Across all leaveone-hospital-out iterations, the most influential predictors consistently included patient age, maximum 30-minute rolling variance, and global sequence standard deviation. These results suggest that prolonged volatility and demographic vulnerability dominate prognostic stratification. Fine-grained micro-temporal patterns, which deep models are theoretically capable of capturing, may not be present in sufficient richness within the BCI abstraction to meaningfully improve discrimination. Thus, the equivalence observed here may reflect signal constraints rather than architectural limitations. Beyond discrimination, this study emphasized probabilistic reliability. Calibration analysis using 15-bin Expected Calibration Error (ECE), Brier score, and reliability diagrams demonstrated modest miscalibration in both modeling paradigms. The Multimodal CNN–BiGRU exhibited a pooled pre-scaling ECE of

approximately 0.146, while XGBoost showed slightly lower miscalibration at approximately 0.122. These differences were small, but they reinforce the importance of evaluating probability quality rather than relying solely on threshold-based metrics. In high-stakes neurocritical care settings, miscalibrated probabilities can undermine clinician trust and decision support integration. Reporting both pooled and mean-of-fold calibration metrics further strengthened institutional interpretability by preventing dominance from higher-volume centers. A particularly important contribution of this work lies in its explicit modeling of predictive uncertainty. Through risk-coverage analysis, we quantified how classification accuracy changes as low-confidence predictions are selectively deferred. At full population coverage, the Multimodal CNN-BiGRU achieved approximately 74% accuracy. By abstaining on low-confidence cases, accuracy increased substantially, surpassing the $\geq$85% threshold when a portion of ambiguous predictions was excluded. This selective prediction paradigm mirrors real-world ICU workflows in which algorithmic systems serve as decision support rather than deterministic arbiters. Rather than forcing binary decisions on uncertain waveforms, the framework allows safe referral to human expertise, thereby mitigating potential harm from overconfident errors. Such uncertainty-aware deployment mechanisms remain underreported in neuroprognostication research yet are critical for responsible integration. While coverage at $\geq$85% accuracy is limited, such a reject-option framework may be valuable in early triage or as a secondary opinion tool rather than as a universal screening mechanism. The incorporation of predictive entropy as a confidence measure enables a clinically meaningful rejection option. By performing Monte Carlo dropout with 50 stochastic forward passes, we obtained a predictive entropy value for each patient, reflecting epistemic uncertainty in the model's posterior estimate. Empirically, entropy values above approximately $H \approx 0.64$ corresponded to predictions that were frequently misclassified under full coverage deployment. Using this entropy threshold as a triage boundary, the model can automatically flag ambiguous cases for expert neurologist review rather than issuing an overconfident binary decision. This selective prediction framework is particularly relevant in neuroprognostication, where false pessimistic predictions may influence withdrawal-of-life-sustaining therapy decisions. Instead of forcing deterministic classification for every patient, the entropy threshold functions as a safety valve: high-confidence predictions can be communicated with greater reliability, while uncertain cases are escalated for multimodal clinical evaluation. In this way, the system transitions from a purely predictive algorithm to a risk-aware decision support tool. The rejection option does not merely improve accuracy numerically; it operationalizes caution in high-stakes environments, aligning algorithmic behavior with ethical and clinical priorities. Fairness analysis was conducted across sex, age strata, and intersectional subgroups. Although minor variations in AUROC and F1 were observed between older and younger cohorts, bootstrap confidence intervals for disparity deltas overlapped zero in primary comparisons, indicating no statistically significant systemic bias. Differences in operating-point metrics largely reflected prevalence imbalances rather than structural model prejudice. Importantly, fairness evaluation extended beyond AUROC to include FPR, FNR, calibration metrics, and subgroup prevalence, providing a multidimensional equity assessment. In multicenter critical care contexts where demographic distributions vary substantially,

Table 4 Comparative landscape of I-CARE/PhysioNet Challenge approaches (official 72h evaluation when reported).

Study	Inputs	Feature/representation	Model	Evaluation	Reported performance
Patidar et al. [15]	EEG + meta	TQWT + scattering; Tucker decomposition	Bagging (DT)	Official (72h)	Score: 0.69 (rank 7)
Venkataramani et al. [21]	EEG+ECG+clinical	PSD + burst-suppression; imputation	XGBoost + RF	Official (72h)	Score: 0.381 (rank 28)
Babu et al. [5]	EEG+ECG+clinical	Hand-crafted PSD/BS/correlation	CNN + LSTM	Hidden val.	Score: 0.63 (EEG-only); ECG 0.28
Kim et al. [12]	Raw EEG + demo	Pretrained ConvNeXt + augmentation	ConvNeXt + MLP	Official (72h)	Score: 0.79 (rank 2)
Rohr et al. [18]	Raw EEG	ResNet embedding; SVM baseline features	ResNet + Transformer	Cross-val	Mean CV score: 0.41; SVM more robust on noisy data
Cauchi et al. [7]	EEG (subset)	Frame-level stats + masking	LSTM	Official (72h)	Score: 0.025 (rank 36/36)
Krones et al. [13]	EEG + clinical	Spectrogram + DenseNet; ROCKET features	Fusion + Ridge/ML	Official (72h)	Submitted: 0.53; local CV AUC: 0.854

such evaluation is essential to ensure safe generalization. The challenge score, defined as TPR at FPR ≤ 0.05, slightly exceeded 0.27 in our evaluation. While this value is lower than those reported by top-performing PhysioNet I-CARE Challenge submissions, interpretation requires contextualization. High-ranking leaderboard entries typically operated directly on raw multichannel EEG waveforms, incorporated spectral and timefrequency features, leveraged larger multimodal physiological inputs, and explicitly optimized architectures toward the low-FPR regime. Table 3 shows the comparative landscape of I- CARE/PhysioNet Challenge approaches, and the methods and performances of some of the participating teams. Note: Official PhysioNet Challenge evaluation uses TPR at fixed FPR ≤ 0.05 at 72h. Some works additionally report cross-validation or hidden-validation results; metrics may not be directly comparable across studies. In contrast, the present study intentionally restricted modeling to the publicly available Dryad BCI representation, a compressed single-channel abstraction of EEG background continuity. The objectives of this work were therefore distinct from leaderboard optimization. Rather than maximizing a single metric, we prioritized multicenter generalizability, statistical equivalence testing, calibration, fairness, and uncertainty characterization. The challenge score is reported transparently for benchmarking but does not represent the primary contribution of this study. The decision to utilize the Dryad BCI dataset instead of raw I-CARE EEG data was driven by considerations of accessibility, reproducibility, and translational feasibility. The raw I-CARE dataset is extremely large and computationally intensive, requiring substantial infrastructure for preprocessing and storage. By leveraging the BCI abstraction, this work demonstrates that clinically meaningful prognostication can be performed using a compact, publicly accessible representation. This approach lowers the barrier to replication and enables deployment in resource-constrained environments without advanced EEG processing pipelines. Future investigations may extend this framework to raw EEG to determine whether richer spectral and spatial representations meaningfully elevate performance beyond the BCI ceiling observed here. An important internal evolution of this work involved transitioning from Trial A to Trial B. The underlying neural architecture remained identical across both trials; improvements stemmed exclusively from systemic refinements. Trial A employed standard binary cross-entropy loss with a fixed 0.5 classification threshold, resulting in a bias toward the majority outcome class. Trial B introduced Focal Loss to emphasize hard-to-classify cases and replaced the arbitrary threshold with a validation-derived Youden index optimization strategy. These changes improved operating-point equilibrium without altering representational topology. This progression underscores that optimization strategy and decision policy can meaningfully influence clinical metrics even when architecture remains constant. Such systemic factors are often underreported but critically shape

real-world performance. Several limitations must be acknowledged. First, the BCI is a compressed abstraction of EEG continuity and does not capture full spectral or spatial dynamics. Second, although multicenter, the cohort size of 607 patients constrains scaling analyses for high-capacity architectures. Third, dichotomization of CPC outcomes may obscure nuanced gradations in neurological recovery. Fourth, the challenge score was not explicitly optimized during training. Finally, while hospitallevel cross-validation approximates institutional deployment, external international validation would further strengthen generalizability. Future research should explore integration of raw EEG features, multimodal physiological fusion, prospective validation studies, and advanced Bayesian uncertainty modeling. Fairnessaware regularization strategies and adaptive calibration mechanisms may further enhance safe deployment. Additionally, extending selective prediction frameworks to dynamic ICU triage scheduling may offer clinically meaningful workflow integration. In summary, this study demonstrates that the BCI representation contains substantial prognostic information and that classical engineered features can approximate deep sequence modeling under signal compression constraints. However, the deep Multimodal CNN-BiGRU provides a flexible probabilistic framework capable of uncertainty quantification, demographic analysis, and systematic deployment evaluation. The equivalence between deep and classical approaches challenges assumptions regarding automatic superiority of raw-sequence modeling and emphasizes the importance of signal richness, calibration, fairness, and operational design in clinical machine learning systems.

6 Conclusion

This study presents a rigorously evaluated, multicenter framework for neurological outcome prediction using 72-hour Brain Continuity Index (BCI) trajectories combined with clinical metadata. Under a hospital-level leave-one-center-out design with strict validation isolation, the proposed Multimodal CNN–BiGRU achieved strong and stable discrimination across institutions while maintaining transparent calibration and uncertainty-aware behavior. Importantly, classical machine learning models trained on 27 engineered features demonstrated statistical non-inferiority relative to the deep sequence model. This finding suggests that within the compressed BCI representation, clinically meaningful prognostic information can be effectively captured using compact, interpretable temporal statistics. The deep architecture therefore provides representational flexibility rather than guaranteed performance gains under signal abstraction constraints. Beyond discrimination, this work integrates calibration analysis, subgroup fairness evaluation, and selective prediction mechanisms within a unified reporting framework. The inclusion of statistical equivalence testing, bootstrap confidence intervals, and explicit hospital-level validation strengthens the reliability of the findings and aligns the modeling pipeline with trustworthy AI principles. The observed challenge score reflects the informational constraints of the compressed BCI representation and the absence of direct multi-channel waveform modeling. In contrast to leaderboard-optimized I-CARE submissions that leveraged raw EEG, multi-sensor inputs, and explicit optimization toward the FPR ≤ 0.05 operating regime, the primary objective of this study was to evaluate reproducible multicenter generalization, calibration,

fairness, and uncertainty under an accessible EEG-derived abstraction rather than to maximize a single competition metric. Overall, the results demonstrate that robust neuroprognostication can be achieved using compressed EEG continuity signals, and that careful evaluation design, statistical rigor, and deployment-aware analysis are essential components of clinically meaningful AI systems.

Declarations

- **Funding:** The authors received no specific funding for this work.
- **Competing interests:** The authors declare no competing interests.
- **Ethics approval:** This study is a secondary analysis of publicly available, de-identified data released by the I-CARE consortium; ethics approval information is provided by the original data contributors.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** The Brain Continuity Index (BCI) dataset is available via Dryad (https://doi.org/10.5061/dryad.2fqz612zv).
- **Code availability:** [29]
- **Author contributions:** **Selay Koyuncu** and **Pardis Azemikhah** contributed equally. Both authors: Conceptualization; Methodology; Formal analysis; Data curation; Project administration; Supervision; Visualization. Pardis Azemikhah: Software; Investigation; Writing – original draft. Selay Koyuncu: Writing – review & editing; Validation. All authors reviewed and approved the final manuscript.

References

[1] Higginson, I.J., Rumble, C., Shipman, C., Koffman, J.: The value of uncertainty in critical illness? an ethnographic study of patterns and conflicts in care and decision-making trajectories. BMC Anesthesiology **16**, 11 (2016) https://doi.org/10.1186/s12871-016-0177-2

[2] James, F.R., Power, N., Laha, S.: Decision-making in intensive care medicine – a review. Journal of the Intensive Care Society **19**(3), 247–258 (2018) https://doi.org/10.1177/1751143717746566

[3] Lighthall, G.K., Vazquez-Guillamet, C.: Understanding decision making in critical care. Clinical Medicine & Research, 156–168 (2015) https://doi.org/10.3121/cmr.2015.1289

[4] Sandroni, C., D'Arrigo, S., Nolan, J.P.: Prognostication after cardiac arrest. Critical Care **22**, 150 (2018) https://doi.org/10.1186/s13054-018-2060-7

[5] Dale, C.M., Sinuff, T., Morrison, L.J., Golan, E., Scales, D.C.: Understanding early decisions to withdraw life-sustaining therapy in cardiac arrest survivors. a qualitative investigation. Annals of the American Thoracic Society **13**(7), 1115–1122 (2016) https://doi.org/10.1513/AnnalsATS.201511-751OC

[6] Robertsen, A., Helseth, E., Laake, J.H., Førde, R.: Neurocritical care physicians' doubt about whether to withdraw life-sustaining treatment the first days after devastating brain injury: an interview study. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine **27**, 81 (2019) https://doi.org/10.1186/s13049-019-0648-9

[7] Admiraal, M.M., Rootselaar, A.-F., Hofmeijer, J., Hoedemaekers, C.W.E., Kaam, C.R., Keijzer, H.M., Putten, M.J.A.M., Schultz, M.J., Horn, J.: Electroencephalographic reactivity as predictor of neurological outcome in postanoxic coma: A multicenter prospective cohort study. Annals of Neurology **86**(1), 17–27 (2019) https://doi.org/10.1002/ana.25507

[8] Azabou, E., Navarro, V., Kubis, N., Gavaret, M., Heming, N., Cariou, A., Annane, D., Lofaso, F., Naccache, L., Sharshar, T.: Value and mechanisms of eeg reactivity in the prognosis of patients with impaired consciousness: a systematic review. Critical Care **22**, 184 (2018) https://doi.org/10.1186/s13054-018-2104-z

[9] Hoedemaekers, C., Hofmeijer, J., Horn, J.: Value of eeg in outcome prediction of hypoxic-ischemic brain injury in the icu: A narrative review. Resuscitation **189**, 109900 (2023) https://doi.org/10.1016/j.resuscitation.2023.109900

[10] Rossetti, A.O., Carrera, E., Oddo, M.: Early eeg correlates of neuronal injury after brain anoxia. Neurology **78**(11), 796–802 (2012) https://doi.org/10.1212/WNL.0b013e318249f6bb

[11] Carrasco-G'omez, M., Keijzer, H.M., Ruijter, B.J., Bru na, R., Tjepkema-Cloostermans, M.C., Hofmeijer, J., Putten, M.J.A.M.: Eeg functional connectivity contributes to outcome prediction of postanoxic coma. Clinical Neurophysiology **132**(6), 1312–1320 (2021) https://doi.org/10.1016/j.clinph.2021.02.011

[12] Amorim, E., Zheng, W.-L., Ghassemi, M.M., Aghaeeaval, M., Kandhare, P., Karukonda, V., Lee, J.W., Herman, S.T., Sivaraju, A., Gaspard, N., Hofmeijer, J., Putten, M.J.A.M., Sameni, R., Reyna, M.A., Clifford, G.D., Westover, M.B.: The international cardiac arrest research consortium electroencephalography database. Critical Care Medicine **51**(12), 1802–1811 (2023) https://doi.org/10.1097/CCM.0000000000006074

[13] Reyna, M.A., Amorim, E., Sameni, R., Weigle, J., Elola, A., Rad, A.B., Seyedi, S., Kwon, H., Zheng, W.-L., Ghassemi, M.M., Putten, M.J.A.M., Hofmeijer, J., Gaspard, N., Sivaraju, A., Herman, S.T., Lee, J.W., Westover, M.B., Clifford, G.D.: Predicting neurological recovery from coma after cardiac arrest: The george b. moody physionet challenge 2023. In: Computing in Cardiology (CinC), vol. 50, pp. 1–4 (2023). https://doi.org/10.22489/CinC.2023.019

[14] Amorim, E., Zheng, W.-L., Lee, J.W., Herman, S., Ghassemi, M., Sivaraju, A., Gaspard, N., Hofmeijer, J., Putten, M.J.A.M., Reyna, M., Clifford, G.,

Westover, M.B.: I-CARE: International Cardiac Arrest REsearch consortium electroencephalography database – Brain Continuity Index. Dryad dataset (2024). https://doi.org/10.5061/dryad.2fqz612zv

[15] Kim, D.-K., Yoon, H.-C., Kim, H.-S., Seo, W.-Y., Kim, S.-H.: Predicting neurological outcome after cardiac arrest using a pretrained model with electroencephalography augmentation. In: Computing in Cardiology (CinC), vol. 50, pp. 1–4 (2023). https://doi.org/10.22489/CinC.2023.077

[16] Venkataramani, V.V., Garg, A., Maity, M., Priyakumar, U.D.: A machine learning approach for outcome prediction in postanoxic coma patients using frequency domain features. In: Computing in Cardiology (CinC), vol. 50, pp. 1–4 (2023). https://doi.org/10.22489/CinC.2023.255

[17] Zhu, M., Xu, M., Gao, M., Yu, R., Bin, G.: Robust eeg characteristics for predicting neurological recovery from coma after cardiac arrest. Sensors **25**(7), 2332 (2025) https://doi.org/10.3390/s25072332

[18] Elmer, J., Chen, J., Turner, A., Shook, B., DiFiore-Sprouse, S., Callaway, C.W., Clermont, G., Dubrawski, A.: How much information is needed to predict outcomes after cardiac arrest? Resuscitation **215**, 110775 (2025) https://doi.org/10.1016/j.resuscitation.2025.110775

[19] Niu, Y., Chen, X., Fan, J., *et al.*: Explainable machine learning model based on eeg, ecg, and clinical features for predicting neurological outcomes in cardiac arrest patient. Scientific Reports **15**, 11498 (2025) https://doi.org/10.1038/s41598-025-93579-0

[20] Zheng, W.-L., Amorim, E., Jing, J., *et al.*: Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks. Resuscitation **169**, 86–94 (2021) https://doi.org/10.1016/j.resuscitation.2021.10.034

[21] Kim, T.J., Suh, J., Park, S.H., *et al.*: System for predicting neurological outcomes following cardiac arrest based on clinical predictors using a machine learning method: The neurological outcomes after cardiac arrest method. Neurocritical Care **42**, 829–838 (2025) https://doi.org/10.1007/s12028-025-02222-3

[22] Krones, F.H., Walker, B., Parsons, G., Lyons, T., Mahdi, A.: Multimodal deep learning approach to predicting neurological recovery from coma after cardiac arrest. arXiv preprint (2024). https://arxiv.org/abs/2403.06027

[23] Rohr, M., Schilke, T., Willems, L., Reich, C., Dill, S., Güney, G., Antink, C.H.: Transformer network with time prior for predicting clinical outcome from eeg of cardiac arrest patients. In: Computing in Cardiology (CinC), vol. 50, pp. 1–4 (2023). https://doi.org/10.22489/CinC.2023.173

[24] Guntunur, S., Varma, S., Saxena, S., Sagere, S., Srinivasa, G.: Analysis of eeg signals for predicting neurological recovery in post-cardiac arrest comatose patients. In: IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC), pp. 64–67 (2024). https://doi.org/10.1109/AIMHC59811.2024.00019

[25] Robertsen, A., Helseth, E., Laake, J.H., Førde, R.: Neurocritical care physicians' doubt about whether to withdraw life-sustaining treatment the first days after devastating brain injury: an interview study. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine **27**, 81 (2019) https://doi.org/10.1186/s13049-019-0648-9

[26] Dale, C.M., Sinuff, T., Morrison, L.J., Golan, E., Scales, D.C.: Understanding early decisions to withdraw life-sustaining therapy in cardiac arrest survivors. a qualitative investigation. Annals of the American Thoracic Society **13**(7), 1115–1122 (2016) https://doi.org/10.1513/AnnalsATS.201511-751OC

[27] Higginson, I.J., Rumble, C., Shipman, C., Koffman, J.: The value of uncertainty in critical illness? an ethnographic study of patterns and conflicts in care and decision-making trajectories. BMC Anesthesiology **16**, 11 (2016) https://doi.org/10.1186/s12871-016-0177-2

[28] Lighthall, G.K., Vazquez-Guillamet, C.: Understanding decision making in critical care. Clinical Medicine & Research, 156–168 (2015) https://doi.org/10.3121/cmr.2015.1289

[29] icare-trustworthy-ai: Uncertainty-aware multicenter neuroprognostication with BCI. GitHub repository. Accessed: 2026-02-28 (2026). https://github.com/paaaraaaz/icare-trustworthy-ai