Predicting Neurological Outcome from Electroencephalogram Dynamics in Comatose Patients after Cardiac Arrest with Deep Learning

Wei-Long Zheng†, Edilberto Amorim†, Jin Jing, Ona Wu, Mohammad Ghassemi, Jong Woo Lee, Adithya Sivaraju, Trudy Pang, Susan T. Herman, Nicolas Gaspard, Barry J. Ruijter, Marleen C. Tjepkema-Cloostermans, Jeannette Hofmeijer, Michel J. A. M. van Putten, and M. Brandon Westover*

***Abstract*—Objective: Most cardiac arrest patients who are successfully resuscitated are initially comatose due to hypoxic-ischemic brain injury. Quantitative electroencephalography (EEG) provides valuable prognostic information. However, prior approaches largely rely on snapshots of the EEG, without taking advantage of temporal information. Methods: We present a recurrent deep neural network with the goal of capturing temporal dynamics from longitudinal EEG data to predict long-term neurological outcomes. We utilized a large international dataset of continuous EEG recordings from 1,038 cardiac arrest patients from seven hospitals in Europe and the US. Poor outcome was defined as a Cerebral Performance Category (CPC) score of 3-5, and good outcome as CPC score 0-2 at 3 to 6-months after cardiac arrest. Model performance is evaluated using 5-fold cross validation. Results: The proposed approach provides predictions which improve over time, beginning from an area under the receiver operating characteristic curve (AUC-ROC) of 0.78 (95% CI: 0.72-0.81) at 12 hours, and reaching 0.88 (95% CI: 0.85-0.91) by 66 h after cardiac arrest. At 66 h, (sensitivity, specificity) points of interest on the ROC curve for predicting poor outcomes were (32,99)%, (55,95)%, and (62,90)%, (99,23)%, (95,47)%, and (90,62)%; whereas for predicting good outcome, the corresponding operating points were (17,99)%, (47,95)%, (62,90)%, (99,19)%, (95,48)%, (70,90)%. Moreover, the model provides predicted probabilities that closely match the observed frequencies of good and poor outcomes (calibration error 0.04). Conclusions and Significance: These findings suggest that accounting for EEG trend information can substantially improve prediction of neurologic outcomes for patients with coma following cardiac arrest.**

***Index Terms*—Cardiac Arrest, Coma, Deep Learning, Electroencephalogram, Outcome Prediction**

This work was supported in part by grants from NIH-NINDS (1K23NS090900, 1R01NS102190, 1R01NS102574, 1R01NS107291, T32HL007901, T90DA22759, T32EB001680), American Heart Association (postdoctoral fellowship), Neurocritical Care Society (NCS research training fellowship), Society of Critical Care Medicine-Weil research grant, Massachusetts Institute of Technology-Philips Clinician Award, and the Dutch Epilepsy Fund (Epilepsiefonds; NEF 14-18). Asterisk indicates corresponding author (M.B. Westover: mwestover@mgh.harvard.edu). W.-L. Zheng and E. Amorim contributed equally as co-first authors.

W.-L. Zheng, J. Jing, and M. B. Westover are with Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. W.-L. Zheng is also with Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA, USA.

E. Amorim is with Department of Neurology, University of California, San Francisco, San Francisco, CA, USA.

O. Wu is with Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

M. Ghassemi is with Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA, and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA.

J. W. Lee is with Department of Neurology, Brigham and Womens Hospital, Boston, MA, USA.

A. Sivaraju is with Department of Neurology, Yale School of Medicine, New Haven, CT, USA.

T. Pang is with Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA.

S. T. Herman is with Barrow Neurological Institute, Phoenix, AZ, USA.

N. Gaspard is with Department of Neurology, Universit Libre de Bruxelles, Brussels, Belgium.

B. J. Ruijter is with Department of Clinical Neurophysiology, University of Twente, Enschede, the Netherlands.

M. C. Tjepkema-Cloostermans and M. J. A. M. van Putten are with Department of Clinical Neurophysiology, University of Twente, Enschede, the Netherlands, and Departments of Neurology and Clinical Neurophysiology, Medisch Spectrum Twente, Enschede, the Netherlands.

J. Hofmeijer is with Department of Neurology, Rijnstate Hospital, Arnhem, the Netherlands, and Department of Clinical Neurophysiology, University of Twente, Enschede, the Netherlands.

I. Introduction

CARDIAC arrest (CA) is the third leading cause of death in the US, with more than 356,000 out-of-hospital cardiac arrests (OHCA) annually [1]. Most patients surviving to hospital admission arrive in coma due to hypoxic-ischemic brain injury, and some patients are treated with targeted temperature management (TTM) to prevent further brain injury [2]. Early and accurate prediction of neurologic outcome is critical for clinical decision making and timely interventions, and several guidelines have been proposed to guide prognostication after cardiac arrest in recent decades. [3], [4] Beyond clinical examination, several ancillary tests can support outcome prediction. These include electroencephalogram (EEG) monitoring, somatosensory evoked potentials, and neuroimaging. [5]–[8] However, there is significant variability between patient presentations and brain injury patterns, making accurate prediction of outcomes challenging.

Recent literature has shown that early EEG patterns observed over the first few days following post-cardiac arrest are strongly associated with good or poor neurologic outcomes, and that the strength of these associations for some features is time-dependent [9], [10]. For example, burst suppression, isoelectric patterns, and certain epileptiform patterns are associated with poor outcomes, with the strength of the association depending on the type and timing, and the strength of this association grows stronger 24 hours or later after cardiac arrest. [9], [11] The association between poor outcomes and burst suppression with identical bursts has been reported to be very strong [12], and isoelectric EEG patterns become strongly predictive of poor outcomes only when these persist 12 hours or later after cardiac arrest. By contrast, a continuous EEG background with normal amplitude within 12 h and preserved EEG reactivity are associated with a high likelihood of favorable outcomes. [7], [9], [12]–[16] However, due to the high volume and heterogeneity of continuous EEG data, clinicians reviewing EEG data manually are unable to provide optimal prognostic information and visual EEG review can suffer from intra- and inter-observer variability [11], [17]–[19]. Thus, despite widespread adoption of EEG monitoring in comatose cardiac arrest patients, full EEG interpretation remains challenging. In contrast, quantitative analysis of continuous EEG offers automated reproducible measurements. [20]–[22]

Although the EEG after cardiac arrest is dynamic, few studies have investigated the prognostic value of EEG trend information. If trends in EEG features carry important prognostic information, algorithms should be able to leverage these trends to make increasingly more accurate predictions with increasing duration of brain activity monitoring. However, previous algorithms have had limited ability to leverage changes across consecutive hours of EEG monitoring. Most recent studies focus on the first 24 hours after cardiac arrest [9], [23], and most prior algorithms make predictions based on isolated time windows within this early period without integrating the evolution of the EEG across time. It is unclear whether long-term EEG dynamics can be leveraged to improve the accuracy of neurologic prognostication, and it is unclear how best to aggregate information across time both within and beyond the first 24 hours.

Recent advances in machine learning (ML) can help deal with the challenges making predictors from complex data in healthcare settings [24]–[28]. ML approaches have been used to leverage EEG data to predict neurological outcomes in comatose patients after cardiac arrest. [21], [29]–[31] However, the performance of some of these algorithms did not improve monotonically with increasing duration of observation, and in fact worsened in one study including data beyond 24 h [32]. While one conclusion could be that EEG beyond the first 24 hours does not add to discrimination between good and poor outcome groups, we hypothesize that prior approaches have not made optimal use of trend information. A previous study demonstrated that a simple time-sensitive model that leverages time-varying features outperforms baseline methods that are time-insensitive when evaluated on the same dataset [31]. More recently, deep neural networks, specifically convolutional neural networks, were shown to perform best in outcome prediction at 12 and 24 hours after cardiac arrest [29]. However, these prior results have an important limitation in that the long-term trends in the EEG are not explicitly modeled. Deep neural networks with the ability to make use of long-term trends in EEG have not yet been explored.

In this study, we develop a deep learning model for neurologic outcome prediction which leverages trend information in continuous EEG data to improve outcome prediction in patients with coma following cardiac arrest. The performance of our proposed model is evaluated on a large multi-center cardiac arrest EEG dataset (1,038 patients), with data from seven hospitals in Europe and the US. We show that performance of the proposed model exceeds that of other prior new models when evaluated in our cohort. [20], [21], [31], [32] Furthermore, we show how our models performance continuously improves with increasing duration of observation, well beyond the initial 24 hours of monitoring.

II. Materials and Methods

A. Dataset

We developed deep learning models using the multi-center cardiac arrest EEG dataset of the International Cardiac Arrest EEG Consortium (ICARE) with 1,038 patients from seven hospitals in Europe and the US (Fig. 1a). The seven hospitals were Medisch Spectrum Twente (Enschede, Netherlands), Rijnstate Hospital (Arnhem, Netherlands), Erasmus Hospital (ULB, Brussels, Belgium), Brigham and Womens Hospital (BWH, Boston MA, USA) , Beth Israel Deaconess Medical Center (BIDMC, Boston, MA, USA), Massachusetts General Hospital (MGH, Boston MA, USA), and Yale New Haven Hospital (YNH, New Haven, CT, USA). The cardiac arrest EEG monitoring protocols at participating institutions were initiated during hypothermia and continued upon rewarming for a total of approximately 48-72 hours. We developed an international multicenter EEG dataset (ICARE, International Cardiac Arrest EEG Consortium), to achieve a large and diverse cohort [29], [31]. The ICARE dataset contains approximately 58,000 hours of prospectively collected clinical EEG data, patient demographic information, and medical information from the time of admission up 6 months after cardiac arrest. The study was based on a retrospective observational cohort. The research protocol was approved by the Institutional Review Boards of participating hospitals. Written informed consent was not required for this retrospective study.

Neurologic outcomes were assessed using the Cerebral Performance Category (CPC) scale (1-5) at 3 or 6 months after hospital discharge after cardiac arrest [8], [33]. Good outcome was defined as a CPC score of 1 or 2 (minimal to moderate neurologic disability), and poor outcome was defined as a CPC score of 3-5 (severe neurologic disability, persistent coma or vegetative state, or death). Four institutions (MGH, BWH, YNH, and BIDMC) assessed best CPC scores retrospectively through chart review at 6 months and one (ULB) at 3 months. In these institutions, CPC scores were not further reviewed for patients who achieved a good outcome (CPC 1-2) or died by hospital discharge [34]. Subjects discharged with a CPC of 3-4 had additional chart reviews performed to evaluate for recovery or worsening in CPC at 6 months from cardiac arrest. Less than 2% of subjects included required this review. Two institutions recorded CPC scores prospectively through phone or in-person interview for surviving patients (Medisch

TABLE I
PATIENT CHARACTERISTICS, GROUPED BY CPC SCORES.

CPC group	CPC 1	CPC 2	CPC 3	CPC 4	CPC 5
Number of patients	303	70	31	17	617
Age (years)	57 (15)	56 (15)	66 (11)	54 (21)	62 (16)
Female gender (%)	29.04	24.29	35.48	47.06	32.25
Shockable rhythm (VFib/VT, %)	71	67	42	41	31
EEG start time (h)	17 (14)	16 (16)	16 (13)	20 (6)	20 (17)
EEG duration (h)	52 (33)	63 (44)	69 (51)	99 (60)	53 (40)
Out-of-hospital CA (N/A)*	232 (21)	50 (6)	17 (4)	14 (0)	439 (43)
TTM (N/A)*	261 (34)	61 (7)	26 (5)	11 (2)	514 (64)

VFib: ventricular fibrillation; VT: ventricular tachycardia; TTM: targeted temperature management; EEG start time (h) is relative to time of cardiac arrest. All numbers related to age and EEG expressed as mean (standard deviation). *For the number of out-of-hospital CA patients and TTM, we didn't have all information available from different hospitals.

Spectrum Twente and Rijnstate Hospital). 665 out of 1038 patients (64%) had a poor outcome. Patient characteristics grouped by CPC scores are summarized in Table I.

The inclusion criteria included non-traumatic cardiac arrest, age$\geq$ 18 years, return of spontaneous circulation (ROSC), Glasgow Coma Scale score $\leq$ 8 on admission, and management with targeted temperature management (TTM). Exclusion criteria were acute cerebral hemorrhage or acute cerebral infarction. The TTM protocol starts as soon as possible after admission to the emergency room or intensive care unit in participating centers with external cooling pads. Goal temperature (32-34 °C or 36 °C) is maintained for 24 hours, and there is gradual rewarming at 0.25-0.5 °C to 37 °C. Neuromuscular blocking agents are used as needed for shivering for all participating centers with exception of the Massachusetts General Hospital, which utilizes neuromuscular blockade continuously throughout TTM. Sedation management during TTM is done at the treating clinicians discretion. Commonly used sedatives and standard dosing ranges are propofol (25-80 mcg/kg/h), midazolam (0.1 mg/kg/h), or fentanyl (25-200 mcg/h). Only one institution (ULB) used midazolam for sedation preferentially, with the remaining institutions using propofol. At participating institutions, recommendations about withdrawal of life-sustaining therapies are a collaborative effort between critical care and neurology teams, following structured protocols. Multimodal neurological prognostication involved serial neurological examinations with a combination of continuous EEG monitoring, head CT or brain MR imaging, neuron specific enolase, and somatosensory evoked potentials as deemed necessary by the treating clinicians.

B. Data Preprocessing and Feature Extraction

EEGs were recorded routinely with 19 electrodes according to the international 10-20 system. Recorded EEGs were heterogeneous across hospitals in terms of channel names, sampling rates, etc. The raw data were standardized by matching channel names, applying digital bandpass filters (0.5-30 Hz), and resampling to 100 Hz. EEGs were re-referenced to 18 bipolar channels (Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, Fz-Cz, Cz-Pz). We chose bipolar referencing for three main reasons: 1) to reduce artifacts such as ECG, which can contaminate the common average reference; 2) because this montage is often found to be useful in clinical practice; and 3) Previous quantitative EEG analysis and modeling in cardiac arrest used bipolar channels [29], [32].

We identified the following typical types of artifacts for each 5-s epoch: 1) abnormally high amplitude values above 500 μV; 2) small standard deviation of the signal ($< 0.2\mu V$) for more than 2 s within the 4 second epoch; 3) overly fast amplitude change with more than 900 μV within 0.1 s; 4) staircase-like spectral patterns (commonly caused by ICU machines such as cooling blankets or pumps). Clinical EEG recorded in the intensive care environment often contains artifacts and noise. Therefore, we developed a preprocessing pipeline to reduce the influence of artifacts and noise. The steps of the pipeline were as follows: 1) an artifact detection algorithm was used to assign an artifact indicator (0/1) to each consecutive 5-s EEG epoch (applied without overlap). 2) Signal quality was calculated as the percentage of clean epochs within each 5-min EEG segment. 3) The quality scores were then used as weights to the EEG features from each segment and the weight averaged features were used as the inputs to the models.

We extracted nine clinically interpretable EEG features for each bipolar channel with a sliding 5-min time window without overlapping: burst suppression ratio, Shannon entropy, δ (0.5-4 Hz), θ (4-7 Hz), α (8-15 Hz), β (16-31 Hz) band power, α/δ ratio, regularity, and spike frequency. The extracted features were averaged over all bipolar channels to provide inputs to the machine learning models. The sequences of EEG features at each 6-h time interval were used as inputs of the time-dependent models. In cases of intermittent missing data (periods when EEG monitoring was interrupted), missing epochs were interpolated to values in the nearest available epochs.

Burst suppression is an EEG pattern consisting of periods of depressed voltage alternating with periods of higher voltage activity. The burst suppression ratio was calculated as the percentage of time in the suppression within a 5-minute interval using a recursive variance estimation approach [35]. Epileptiform discharge detection was performed using an automated detection algorithm, SpikeNet, described in our previous work, and epileptiform discharge frequency (number of discharges / 5 mins) was the feature utilized to represent epileptiform discharges in our model. [36] Shannon entropy measures signal complexity. Regularity is a measure used

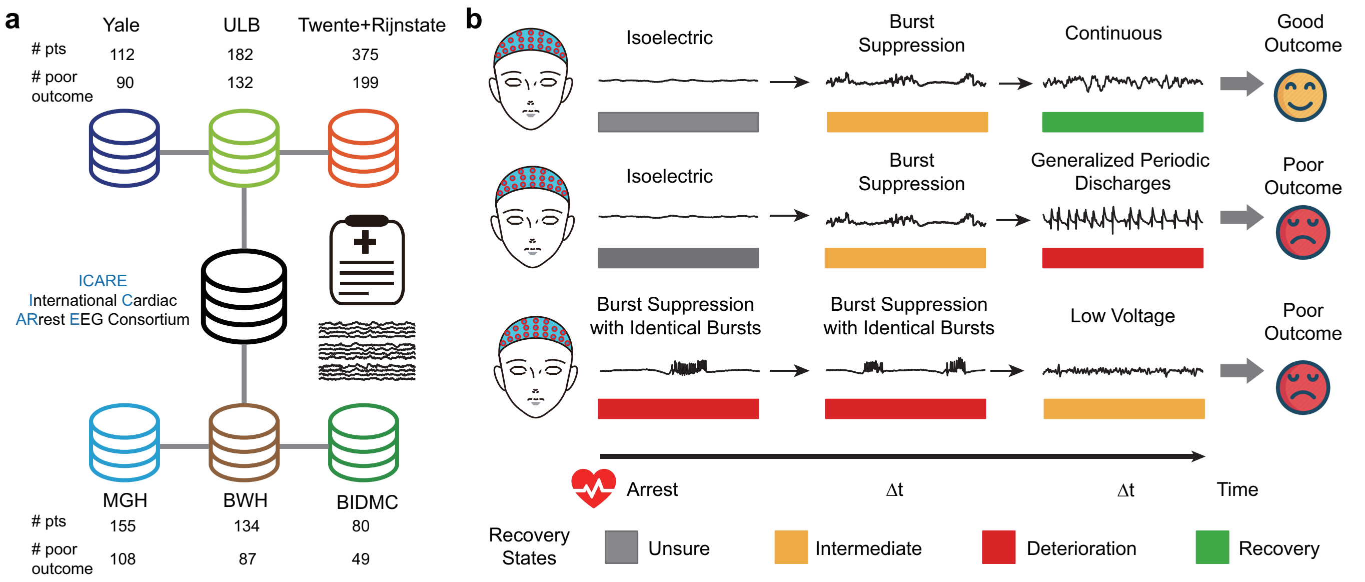

Fig. 1. Study framework. a, We used a large cardiac arrest EEG dataset (ICARE) from seven university-affiliated hospitals in Europe and the US to develop and externally validate the generalization of our prediction models across centers. b, Illustration of the importance of evolution over time with associations between EEG patterns and outcome after cardiac arrest. For example, a rapid transition from an isoelectric state to burst suppression to continuous activity within 12 hours after cardiac arrest usually portends good outcome. EEG patterns present at any given time might not consistently differentiate outcomes. Both the occurrence and the temporal dynamics of EEG patterns contribute to optimally predicting neurologic outcome.

in prior work to separate burst suppression patterns from continuous patterns [20]. For calculating regularity, the EEG signal was smoothed with a moving average, and the data points of smoothed signals were sorted in descending order [20]. The normalized standard deviation of the sorted signal was calculated was a feature for regularity. δ, θ, α, β band power, and α/δ ratio were calculated using the short time Fourier transform with Hamming windows.

C. Model Architecture

Our approach views neurologic outcome prediction as a progressive goal, based on analysis of the evolution of brain states. The states are manifest by different characteristic EEG patterns (Fig. 1b). Prior work by us and others [31], [37] shows that some EEG patterns are strongly associated with a good or poor outcome when seen at any time, e.g., epileptiform patterns (e.g. generalized periodic discharges on a flat background or burst suppression with identical bursts), while the prognostic significance of some intermediate EEG patterns is strongly time dependent, e.g., discontinuity in the EEG [31]. We aimed to endow our outcome prediction model with the ability to capture long-term EEG dynamics to improve overall prediction performance. To achieve this, we developed a time-dependent deep learning model with bidirectional long-short time memory recurrent neural networks (Bi-LSTM).

The input sequences for this model have two components that are concatenated: 1) a mean historical feature sequence: this is obtained by averaging the sequences of feature vectors from all prior 6-hour epochs. Each such sequence contains 72 feature vectors (one from each consecutive 5-minute window), and averaging these sequences produces a single average sequence. Epochs with missing data were interpolated to values in the nearest available epochs prior to averaging. The dimensions of this average sequence are 9 × 72 (9 features in each feature vector × 72 consecutive 5-minute periods in the 6-hour epoch). This average sequence of feature vectors provides historical context for the network in which to evaluate data from the current 6-hour window. 2) A current sequence: the sequence of EEG feature vectors from the current 6-h window. The dimensions of this sequence are also 9 × 72 (9 features in each vector × 72 consecutive 5-minute periods). This arrangement is illustrated in Fig. 2b.

The Bi-LSTM learns temporal dependencies between time steps in the EEG time series by forward and backward processing (Fig. 2a). LSTM introduces multiple gating mechanisms to address the vanishing gradient problem in the backpropagation through time algorithm. The hyperparameters of the neural network were tuned by cross-validation. The best network architecture consisted of four Bi-LSTM layers, three dropout layers, one fully connected layer, and a softmax layer (Fig. 2c). We used multilayered Bi-LSTMs, which mapped the input time series into multiple hidden features. The last element of the output sequence from the top-level Bi-LSTM layer was used as the input for a fully connected layer. Dropout was used during training to help avoid overfitting, and a softmax layer was used to calculate the posterior probability of neurologic outcome. Cross entropy was used as the loss function. Stochastic gradient descent with momentum (SGDM) optimizer was applied to train the deep neural networks. Training samples for the neural network consisted of 6-h EEG time blocks. The final stage of the neural network operating on each 6-hour block (NOPM, neurologic outcome prediction module) produces an estimate of the probability that the final neurologic outcome will be poor. In order to leverage information in past EEG time windows, we developed a sequence of Bi-LSTMs and averaged the output probabilities to arrive at the current predicted probability of a poor outcome (Fig. 2d).

D. Baseline Comparison

We compared the performance of our proposed model with state-of-the-art models on the same dataset. Previous studies found that a simpler convolutional architecture sometimes outperforms canonical recurrent networks, e.g., LSTM [38]. A recent study applied convolutional neural networks to outcome prediction and achieved better performance than previously reported predictors [29]. Therefore, we included a convolutional architecture called temporal convolutional network (TCN) for comparison [38]. TCN performs dilated causal convolution using multiple stacked convolutional layers. With dilated convolution, higher level convolutional layers have larger receptive fields. The TCN architecture also consists of multiple residual blocks, which allows layers to learn modifications to the identity mapping. [38] Another time-dependent model called a sequence of generalized linear models with Elastic Net regularization (SGLM with Elastic Net) was proposed recently [31]. This approach allows models operating at later time points later to consider both past and present features when making predictions. SGLM with Elastic Net can automatically select features based on $\ell 1$ and $\ell 2$ normalization. A conventional baseline classifier, Random Forest, was evaluated to show the performance of models without time dependency.

E. Hyperparameter Tuning

For Bi-LSTMs, we tuned the following hyperparameters: number of layers, number of neurons in each layer, maximal epochs. The ranges of numbers of layers and neurons were [1, 2, 3, 4] and [10, 20, 30, 40, 50], respectively. The maximal epochs were tuned in the range [50 100]. Training data were shuffled every epoch and early stopping was used. We used internal cross validation for hyperparameter tuning (training and validation sets). The best hyperparameters were determined based on the average performance in internal cross validation using an validation set (a subset of the training data). The hyperparameters in each fold were the same in internal cross validation. For TCNs, four residual blocks were used containing dilated causal convolution layers with each 170 filters of size 15. The number of filters was tuned in the range [150, 250] with a step of 10. Filter size was tuned in the range [3, 15] with a step size (stride) of 2. The penalty parameter of SGLM with Elastic Net was tuned with the values of 0.5 and 1. For Random Forest, the number of trees was tuned between 20 and 90 with a step of 10. The best penalty parameter α in SGLM with Elastic Net was 1 and the best number of trees in Random Forest was 60.

F. Performance Evaluation Metrics

To quantify the stability of model performance, we used 5-fold external cross validation and report average performance and 95% confidence intervals. We randomly partitioned available data into 5 folds, where 4 folds were used train model parameters (training and validation sets in internal cross validation) and the remaining 1-fold was used for model evaluation (test set). The split of training, validation, and test sets was patient-independent within each of the 5 folds. Data from the same patients were exclusively in either in the training set or test set; no patient ever had data in both sets. The area under the receiver operating characteristic curve (AUC-ROC) and calibration error were used as evaluation metrics. Calibration error compares predicted probabilities with the observed event frequencies. The averages over five folds were calculated for comparison. The 95% confidence intervals were calculated using the approach of Hanley and McNeil [39], [40]. Statistical significance was evaluated using t-test and p values below 0.05 were considered as statistical significance. We compared the sensitivity and specificity with a thresholded score from the models (99%, 95%, and 90%).

Due to patient privacy in multiple hospitals, the data in the study are not available to the publicity. The processing pipeline and model implementations were based on standard model libraries and scripts in Python and MATLAB. The statistical analysis code used in the study is available from the corresponding author on reasonable request.

III. Results

A. Performance Evaluation

We compute all performance measures for each 6-h time interval between 12-96 h after cardiac arrest. To quantify the stability of these performance measures, we perform 5-fold cross validation. The reported AUC-ROC and calibration errors are averages over the 5-folds, with accompanying confidence intervals and standard deviations. We compared performance of several state-of-the-art time dependent models (Temporal Convolutional Network (TCN), Sequence of Generalized Linear Models (SGLM) with Elastic net regularization) and the baseline model (Random Forest).

Sequences of Bi-LSTMs outperformed the other models (Fig. 3a). Sequences of Bi-LSTMs, Sequences of TCNs, and SGLM with Elastic net were all able to leverage long term temporal dependencies to improve predictions. The performance of these three models increased approximately monotonically with time. The other two models with short-term time dependencies (independent Bi-LSTMs and TCNs) and Random Forest achieved better performance in two time-ranges: approximately 24-42 h and 66-78 h after cardiac arrest. The look-back strategy implemented in the Bi-LSTMs model was able to effectively leverage historical predictions and provide a trajectory of outcome risk for individual patients.

Performance of the various models was similar early after cardiac arrest (before 18 h), while performance of the Bi-LSTM model moderately increased to 0.87 (95% confidence interval, 95% CI: 0.84-0.89, standard deviation, std: 0.03) at 42 h and reached its maximum value of 0.88 (95% CI: 0.85-0.91, std: 0.03) at 66 h. The AUC improvement of the sequence of Bi-LSTM model at 66h compared to Bi-LSTM, sequences of TCNs, TCN, SGLM with Elastic net, and Random Forest was 0.03*, 0.02, 0.08*, 0.02, and 0.07*; where '* indicates passing a test of statistical significance ($p < 0.05$, t-test).

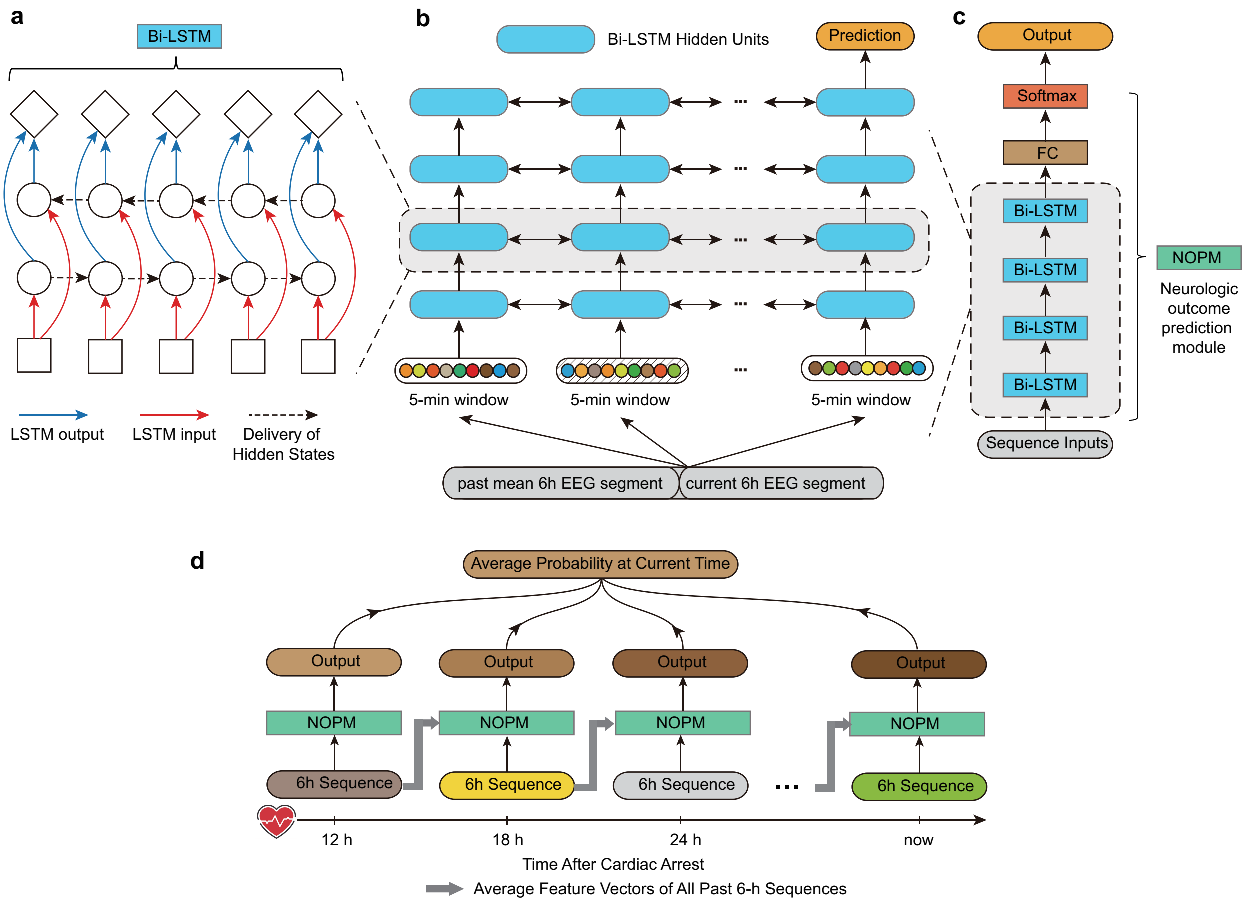

Fig. 2. Model architecture of a sequence of Bi-LSTMs. a, Dependencies between time steps in the EEG sequences were learned by a Bidirectional LSTM. b, A time-dependent deep learning model was developed that takes as input 6-h sequences of past mean and current EEG feature values. The outputs of hidden states in the last Bi-LSTM block were used for prediction. In cases of intermittent missing data (periods when EEG monitoring was interrupted), missing epochs (shaded blocks) were interpolated to values in the nearest available epochs. c, The best network architecture of individual 6-h time blocks consists of four Bi-LSTM layers, three dropout layers, one fully connected layer, and one softmax layer. The neural network was called a neurologic outcome prediction module (NOPM). d, To leverage the output probabilities of Bi-LSTMs at different time blocks and obtain more stable and robust predictions, we averaged the output probabilities of a sequence of Bi-LSTMs until now as the final prediction probabilities.

Although predictions made by the model are probabilities, it is customary to compare these to thresholds and report the statistical performance of the resulting binary predictions. Doing this, performance of the model at 66 h was as follows. For predicting poor outcomes, at specificity thresholds of 99%, 95%, and 90%, the models sensitivity was 32%, 55%, and 62%, respectively; whereas at sensitivity thresholds of 99%, 95%, and 90%, specificity was 23%, 47%, and 62%. For predicting good outcomes, at specificity thresholds of 99%, 95%, and 90%, sensitivity was 17%, 47%, and 62%; whereas at sensitivity thresholds of 99%, 95%, and 90%, specificity was 19%, 48%, and 70%.

The improvement of all models with increasing time provides evidence that leveraging long-term time dynamics of EEG signals provides improved ability to predict neurologic outcome. Sequences of Bi-LSTMs, Sequences of TCNs, and SGLM with Elastic net had consistent improvement in performance with more observations (from mean AUC of 0.78, 0.77, and 0.75 at 12 h to mean AUC of 0.88, 0.86, and 0.87 at 66 h, respectively). The improvement of the three models was statistically significant ($p < 0.01$, t-test).

It should be noted that the numbers of patients with available EEG data varied over time (Fig. 3b). The numbers of patients increased initially and decreased later, reaching a maximal value of 826 during the time period 24-30 h. The ROC curves at different times are shown in Fig. 3c.

B. Calibration Risk

Model calibration was evaluated by comparing the predicted probability of a poor outcome with the proportion of patients who had a poor outcome. We compared calibration curves at different time intervals and calculated calibration errors to quantify performance (Fig. 3d). Calibration error was defined as the absolute deviation from the diagonal line, which represents perfect calibration (lack of systematic errors of over- or under-prediction). Model calibration improved from 12 h to 60 h and deteriorated after 60 h. Calibration error at 66 h was 0.04. Our proposed model was well calibrated, with good agreement between the observed proportions of poor outcomes and predicted probabilities of poor outcomes.

C. Subgroup Analysis

Having investigated overall performance on the whole cohort, we next investigated prediction performance in individual

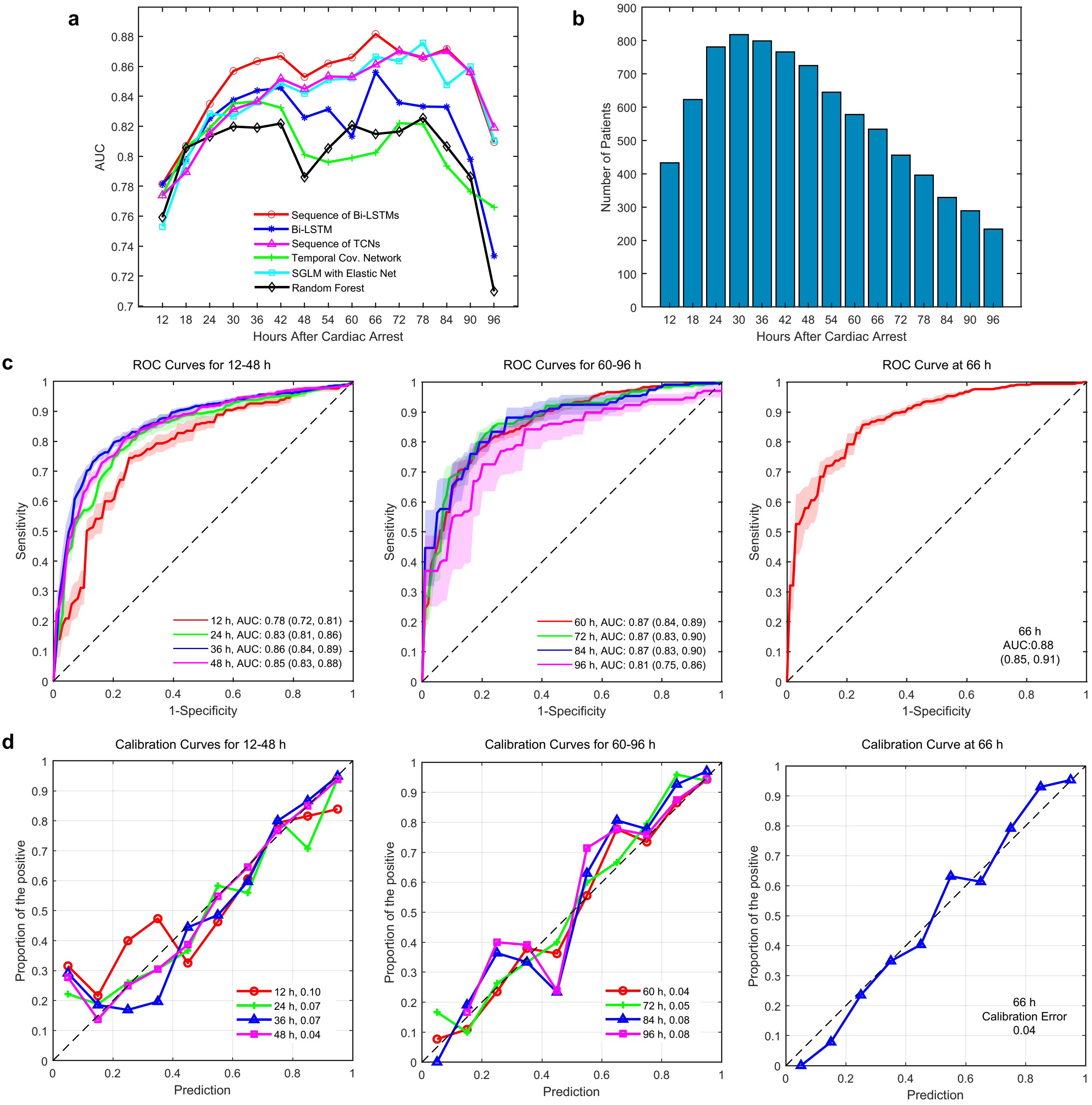

Fig. 3. Model performance of different models in outcome prediction. a, Mean AUC values of different models within each 6-hour time interval. Sequences of Bi-LSTMs (red line) performed best, exhibiting consistent improvement in performance with more observations (from mean AUC of 0.7814 at 12 h to mean AUC of 0.8815 at 66 h). b, Numbers of patients with EEG available with respect to time after cardiac arrest. c, Mean ROC curves at different time intervals (12-48 h, 60-96 h, and 66 h). Shaded areas indicate the standard errors in 5-fold cross validation. d, Calibration curves at different time intervals (12-48 h, 60-96 h, and 66 h). The numbers are calibration errors (deviations from the diagonals).

patients and CPC groups. Fig. 4a makes evident qualitatively (in a colormap) that the sequence of Bi-LSTM prediction probabilities over time in all individual patients. For some patients identified initially as having a low predicted probability of a good outcome, the predicted probability of a good outcome increases progressively as additional observations come in over time, and in general, predicted probabilities are more accurate at later time points. These results support our starting hypothesis, that leveraging long term EEG dynamics can improve prediction performance of neurologic prognostication models. While model predictions generally agree well with observed outcomes, in keeping with the probabilistic framework, the models predictions are not infallible. Poor outcomes occasionally occur despite confident predictions of a good outcome, and vice versa (Fig. 4a).

Next, we grouped outcome prediction probabilities by CPC scores (Fig. 4b). The mean predicted probability of poor outcome within each CPC group was consistent with the

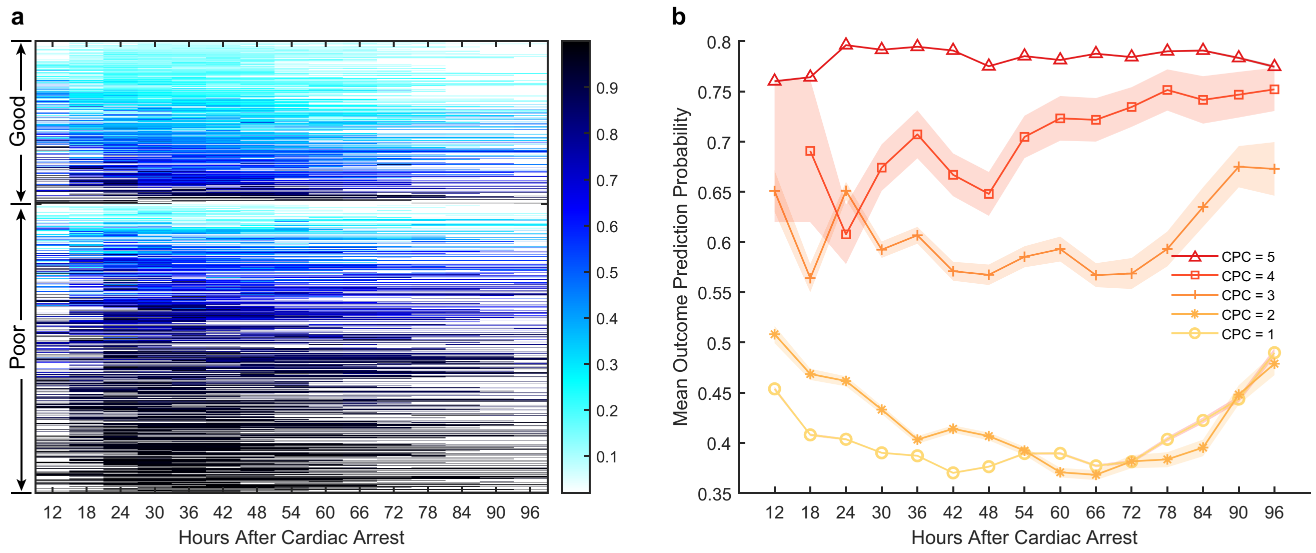

Fig. 4. Prediction probabilities of poor outcome over time for individual patients and individual CPC groups. a, Individual prediction probabilities of poor outcome can change over time. Each row shows the output probabilities from our model for one patient over consecutive 6h blocks, the darker the color, the higher the predicted probability of poor outcome. Patients in each outcome group are sorted based on the mean prediction probabilities. Generally, the group with poor outcomes has substantially higher predicted probabilities of poor outcomes. b, Predicted probabilities over time, grouped by final CPC scores. A CPC score of 1 denotes good recovery while CPC score of 5 denotes death. The overall mean predicted probabilities were consistent with the expected order of CPC scores.

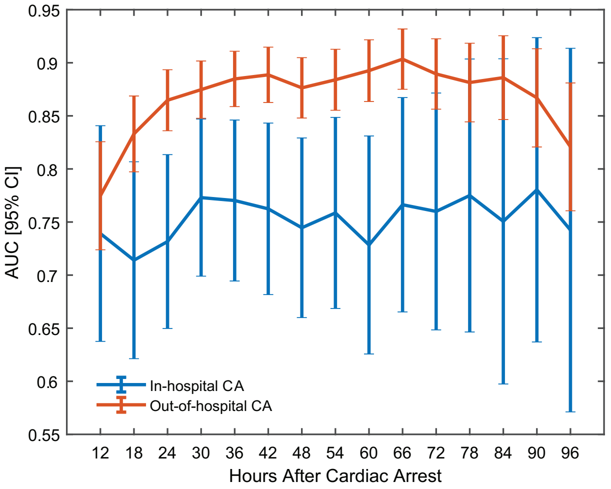

Fig. 5. Model performance of out-of-hospital cardiac arrest and in-hospital cardiac arrest over time.

ordinal ordering of CPC scores. The CPC 5 group had the highest mean prediction probabilities, while the CPC 1 group had the lowest mean prediction probabilities of poor outcomes. The mean prediction probabilities of the CPC 1-5 groups were 0.41, 0.42, 0.61, 0.71, and 0.78, respectively. There was a large probability gap between the group with good outcomes (CPC 1-2) and the group with poor outcomes (CPC 3-5).

For subgroup analysis, we evaluated the model performance for patients with out-of-hospital cardiac arrest and in-hospital cardiac arrest, respectively, after five-fold cross validation (Fig. 5). Most patients in our dataset were patients with out-of-hospital cardiac arrest: 761 patients (73%) had out-of-hospital cardiac arrest, 203 (20%) in-hospital cardiac arrest, and 74 patients (7%) did not have that data available. Overall, the performance of the out-of-hospital cohort were better than those of the in-hospital cohort. The performance of the in-hospital cohort were similar over the time intervals after CA while those of out-hospital cohort increased moderately over time and reached a best AUC of 90% [88%, 93%] at 66 hours after CA.

We last investigated whether model performance varied across patients cared for at different institutions. Model performance varied between institutions (Table II). Notably, outcomes were most predictable early (0-24 hours) on within the two Dutch hospitals (UTW, RIJ), reaching an AUC of 89% by 24 hours, whereas outcome predictability reached only AUC of 66% within the Belgian hospital (ULB).

D. Visualization

The modeling framework was inspired by actual clinical decision making, which considers current EEG information in context with historical information to predict neurologic outcome. Model performance on five typical cases is illustrated in Fig. 6, each with a different CPC score. The mean spectrograms and corresponding EEG snapshots are shown. From the figure, we see that the prediction probabilities follow the rank order of neurologic outcomes (CPC scores).

The first two patients with good outcomes (CPC 1 and 2) had continuous EEG patterns with normal amplitudes during recovery. Early improvements to continuous EEG patterns usually indicate a good outcome. Their spectrograms demonstrate improving power in low frequency bands. Prediction probabilities of poor outcomes were consistently low for both patients over time. The patient with CPC 3 had isoelectric EEG early at 12-24 h after cardiac arrest. Early isoelectric EEG had an intermediate probability of a poor outcome. However, prolongation of the isoelectric pattern increased the

TABLE II
MODEL PERFORMANCE FOR INDIVIDUAL INSTITUTIONS (AUC, 95% CONFIDENCE INTERVALS)

Time Interval	12 h	18 h	24 h	30 h	36 h
BIDMC	0.68 [0.50,0.85]	0.74 [0.61,0.87]	0.80 [0.69,0.90]	0.81 [0.72,0.91]	0.83 [0.75,0.92]
BWH	0.82 [0.69,0.95]	0.75 [0.63,0.86]	0.73 [0.63,0.82]	0.76 [0.67,0.85]	0.71 [0.62,0.80]
MGH	0.77 [0.59,0.94]	0.80 [0.70,0.91]	0.88 [0.81,0.95]	0.90 [0.84,0.96]	0.88 [0.81,0.94]
ULB	0.62 [0.48,0.76]	0.64 [0.52,0.76]	0.66 [0.56,0.76]	0.71 [0.61,0.80]	0.75 [0.66,0.85]
UTW+RS	0.82 [0.76,0.87]	0.85 [0.81,0.90]	0.89 [0.85,0.92]	0.89 [0.86,0.93]	0.90 [0.87,0.94]
YNH	0.59 [0.41,0.77]	0.81 [0.70,0.92]	0.87 [0.78,0.95]	0.90 [0.83,0.97]	0.93 [0.87,0.98]
Time Interval	42 h	48 h	54 h	60 h	66 h
BIDMC	0.82 [0.73,0.92]	0.82 [0.72,0.91]	0.85 [0.76,0.94]	0.85 [0.76,0.94]	0.85 [0.74,0.95]
BWH	0.74 [0.65,0.83]	0.73 [0.64,0.83]	0.76 [0.67,0.85]	0.73 [0.64,0.83]	0.73 [0.63,0.83]
MGH	0.88 [0.82,0.94]	0.84 [0.77,0.91]	0.87 [0.80,0.93]	0.89 [0.83,0.95]	0.93 [0.88,0.98]
ULB	0.77 [0.67,0.87]	0.69 [0.56,0.81]	0.63 [0.46,0.80]	0.61 [0.42,0.80]	0.66 [0.47,0.84]
UTW+RS	0.91 [0.87,0.94]	0.91 [0.87,0.94]	0.91 [0.87,0.95]	0.91 [0.87,0.95]	0.91 [0.87,0.95]
YNH	0.92 [0.86,0.98]	0.91 [0.85,0.97]	0.91 [0.85,0.98]	0.94 [0.88,0.99]	0.95 [0.89,1.00]
Time Interval	72 h	78 h	84 h	90 h	96 h
BIDMC	0.86 [0.76,0.96]	0.85 [0.73,0.96]	0.89 [0.78,1.00]	0.88 [0.76,1.00]	0.86 [0.71,1.00]
BWH	0.71 [0.59,0.83]	0.77 [0.66,0.89]	0.81 [0.70,0.92]	0.79 [0.67,0.92]	0.70 [0.54,0.86]
MGH	0.90 [0.83,0.96]	0.92 [0.86,0.98]	0.92 [0.86,0.99]	0.93 [0.86,1.00]	0.92 [0.84,1.00]
ULB	0.63 [0.42,0.83]	0.60 [0.40,0.80]	0.56 [0.34,0.78]	0.54 [0.32,0.76]	0.46 [0.22,0.70]
UTW+RS	0.91 [0.86,0.95]	0.91 [0.86,0.96]	0.89 [0.83,0.95]	0.88 [0.81,0.95]	0.85 [0.76,0.94]
YNH	0.94 [0.88,1.00]	0.90 [0.80,0.99]	0.95 [0.87,1.00]	0.94 [0.87,1.00]	0.94 [0.85,1.00]

BIDMC: Beth Israel Deaconess Medical Center, BWH: Brigham and Womens Hospital, MGH: Massachusetts General Hospital, ULB: Erasmus Hospital, Universit Libre de Bruxelles, UTW: Medisch Spectrum Twente, and Rijnstate Hospital, University of Twente, YNH: Yale New Haven Hospital.

probability of a poor outcome (from 32.77% at 12 h to 48.03% at 24 h). Later the EEG evolved to have more and more epileptiform discharges (generalized periodic discharges) and the patient experienced seizures. With continuation of unfavorable EEG patterns throughout the first 72 hours, the prediction probability of a poor outcome from our model reached 81.56% by 72 hours. The patient with CPC 4 had a high burst-suppression ratio with epileptiform discharges lasting for more than 12 hours. The evolution of the EEG to continuous patterns occurred late (e.g., after 48 h). Therefore, the output probabilities of a poor outcome were relatively high over time. The EEG of the patient with the worst outcome (CPC 5) showed persistent voltage suppression (last row of Fig. 6). This patient had a high burst-suppression ratio with highly epileptiform bursts. The prediction probabilities for this patient were high throughout the entire course of EEG monitoring (over 95%).

IV. DISCUSSION

Our results demonstrate that a deep learning model that leverages EEG dynamics can provide accurate neurologic outcome predictions post-cardiac arrest that become more accurate as time passes. Our time-sensitive models accuracy continued to increase as additional EEG data was included, reaching maximum predictive accuracy at 66 hours (AUC 0.88). The model was well calibrated, with observed proportions of poor outcomes closely matching predicted probabilities. Further, outcome probabilities mapped closely onto observed outcome categories, following the rank order of CPC scores. These individual-level outcome probabilities of the model are suitable for risk stratification for neurologic outcome prediction after cardiac arrest. Additional relevant features of this study is its size, with more than 1,000 prospectively collected cases, and the inclusion of patients from seven different hospitals from three countries (United States, Netherlands, and Belgium).

This work builds on several prior studies using quantitative analysis of EEG data to predict neurologic outcome in postanoxic coma. Most prior studies have used time-insensitive models, which make predictions based on EEG data available from specific epochs, e.g. 0-12 hours, 12-24 hours. [20], [21], [29], [31], [32] Partial exceptions are the Cerebral Recovery Index (CRI) models, of which there have been three versions [20], [21], [32], all from studies performed in the Netherlands. The first utilized 109 patients from 1 hospital; the second, 238 patients from two hospitals (UTW, RS); the third, from 551 patients from the same two hospitals (a subset of the I-CARE cohort in the present study). Unlike most prior work, the three CRI studies investigate prediction performance over time. Maximal AUC was achieved in the original CRI paper at 18 h (0.94) using a hand-crafted parametric model with 5 QEEG features [20]; at 12 h (0.92) in the second CRI using random forest model employing 9 QEEG features [21]; and at 12 h (0.94) in the third CRI employing 44 features in a random forest model, supplemented with a feature selection procedure (LASSO regression) [32]. In contrast to the present work, none of the three CRI models attempted to leverage temporal trends to improve prediction performance over time.

In more recent work [29], the CRI authors utilized data from 895 patients from 5 Dutch hospitals, to train a convolutional neural network (CNN) to predict neurologic outcome at two time points (12 and 24 hours). The authors also tried concatenating EEG inputs from 12 and 24 hours. Maximal performance on the validation set was achieved at 12 hours (AUC 0.92); though performance for the model that combined information from 12 and 24 hour was essentially the same (AUC 0.91). However, the authors did not explicitly investigate the prognostic value of EEG trends and did not attempt to leverage the full temporal evolution of the EEG; it is possible that even better performance could have been achieved by

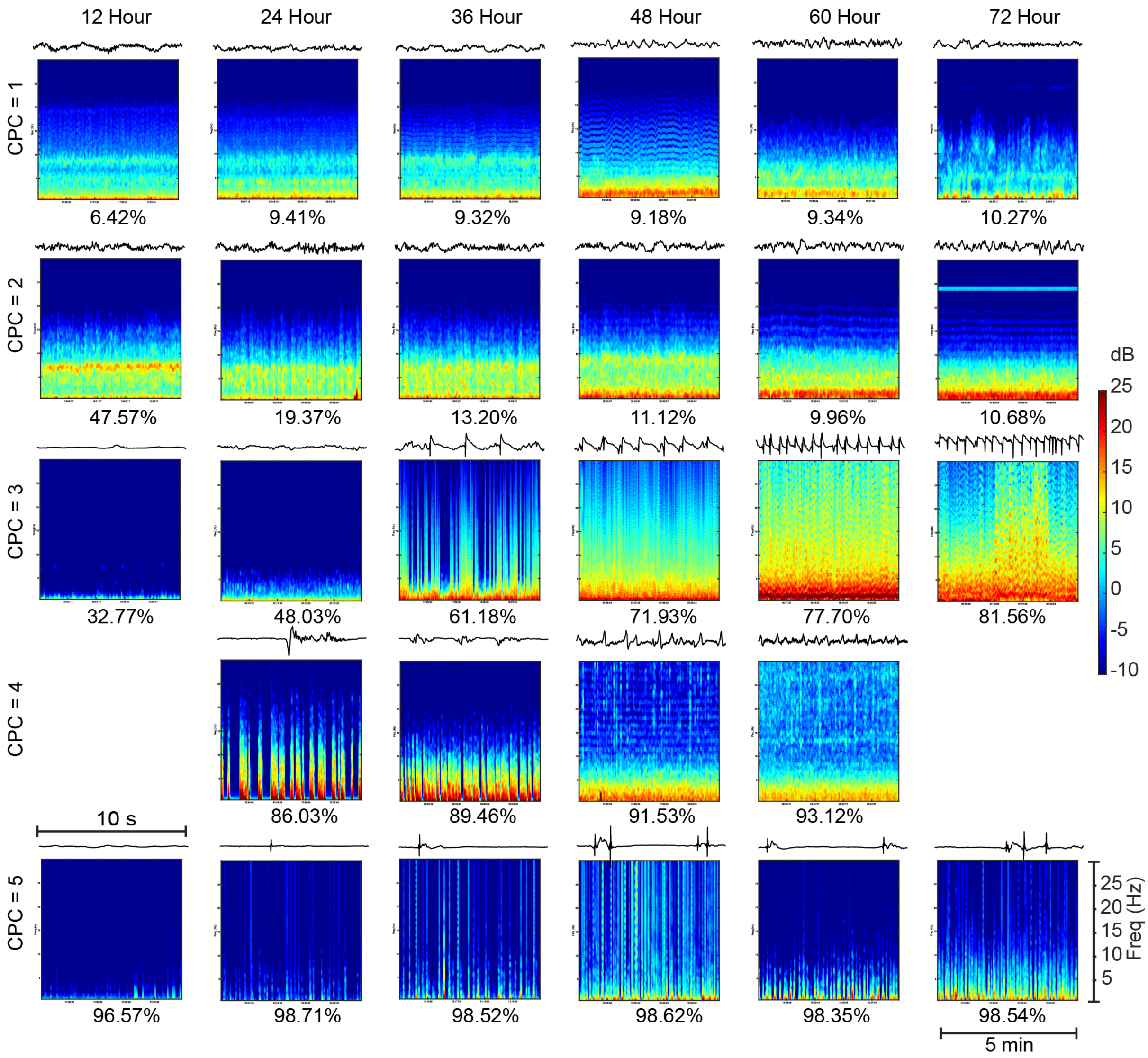

Fig. 6. Model performance on sample patients. Each row illustrates the mean multi-taper spectrogram and EEG waveforms in multiple time blocks. At the bottom of each spectrogram, prediction probabilities of the model for the corresponding EEG segments are shown. The time length of EEG snapshots was 10 s while the spectrograms spanning a 5-min time window are shown. Generally, continuous EEGs had low prediction probabilities of poor outcomes while burst-suppression patterns and epileptiform discharges produced high prediction probabilities of poor outcomes.

leveraging temporal trends. One recent study that did explicitly attempt to construct a time-sensitive model utilized data from 438 patients from four US hospitals, to train a sequence of Generalized Linear Models (SGLM) with 52 QEEG features as input (with elastic net feature selection). [31] The time-sensitive SGLM model demonstrated monotonically improving prediction over time, by making use of a memory bank of progressively more EEG feature vectors from prior epochs, and achieved a maximal AUC of 0.83 by 72 hours. The predictive performance of individual features recorded at different time points changed over time, indicating that the discriminative power of EEG data is both time-dependent and feature-specific. On the same data set, the time-sensitive model performed better than a random forest model based on the second CRI model (AUC 0.83 vs. 0.74, respectively). In addition to measuring performance by AUC, the SGLM paper also introduced the concept of model calibration (how well the predicted probability of good or poor outcome agrees with the observed frequency of outcomes) as a key indicator of model performance, arguing that such probabilistic information is more relevant to clinical decision making than simple binary predictions (with accompanying measures of sensitivity and specificity). The SGLM model was shown to have excellent calibration across the initial 72 hours of EEG monitoring, superior to several time-insensitive approaches [31].

In the current study, we utilized data from 1038 patients from 7 hospitals in 3 countries, the largest and most diverse dataset assembled to date to develop machine learning models to predict neurologic outcome in postanoxic coma. We directly compared a wide variety machine learning models on the same data, including several of the prior best performing models (e.g. random forest and SGLM), in addition to several new model types. Best performance was achieved by a time-sensitive model Bi-LSTM model, which showed monotonically increasing performance up to 66 hours. The Bi-LSTM model performed slightly better than the time-sensitive elastic net model (AUC 0.88 vs. 0.86, respectively). In addition, the Bi-LSTM model was superior to other state-of-the-art machine learning models (sequence of TCN and Random Forests).

It is important to note that model prediction statistics (e.g. AUC values) cannot be directly compared across prior studies. Important differences between studies include: 1) The current data set is larger; 2) the current data set is more heterogeneous, coming from seven hospitals and three different countries. Indeed, our data suggest that predictability of neurologic outcome likely varies substantially between centers, thus between-center heterogeneity may be consequential. Possible reasons for differential predictability include differences in patient characteristics, care practices, and decision-making regarding withdrawal of care. Careful future study of this issue is warranted. 3) Model training and validation strategies differ across studies. 4) Model evaluation practices differed across studies. An important feature of the original elastic net study and the current Bi-LSTM model lacking in prior studies is the emphasis on model calibration. Calibration provides a measure of a models ability to provide clinically relevant probabilistic estimates of risk, which can be done at the individual patient-level and across all predicted probabilities, without artificially imposing pre-specified binary thresholds.

Our study has several important limitations. 1), As seen in Fig. 4a, prediction is not perfect; there exist cases where model fails to make the correct prediction consistently throughout EEG monitoring. It is possible that calibrating the general-purpose model developed herein to characteristics of individual patients could further improve prediction performance. 2), Our model utilized only EEG information. Baseline patient and treatment characteristics are also associated with outcome after cardiac arrest, e.g., location of arrest, first recorded rhythm, time from 911 call to sustained restoration of circulation, and method of induced hypothermia/targeted temperature management. Incorporating a wider array of information might further improve outcome predictability. However, not all of these clinical variables were available due to different data collection protocols in different centers. 3), In the present study we focused on nine clinically interpretable EEG features. We did not include all features known to be associated with poor outcomes. For example, as mentioned above, we did not quantify similarity between bursts in burst suppression. Similarly, although we included information about the frequency of epileptiform discharges and background amplitude, we did not explicitly account for the periodicity of discharges, nor did we explicitly construct a feature which looked for the conjunction of generalized periodic discharges and a flat or low voltage background, another pattern strongly associated with poor outcomes. [9], [11] It is possible that including information about such features more explicitly would further improve model performance. 4), It is possible that additional 'data driven' features, beyond those described in the literature, might further improve model performance. Some prior EEG studies (outside the field of cardiac arrest prognostication) have developed hybrid deep neural networks which combine convolutional neural networks (CNN) and recurrent neural networks (RNN) for EEG time series, where EEG features are automatically learned from raw waveforms with CNN and time dependencies between are modeled with RNN models. Such hybrid network architectures (CNN-RNN) have been validated in some time series applications [41], [42], and this is a promising future direction for our work. 5), Treating physicians were not blinded to EEG results in the present study, and thus may have used these results for decision making regarding continuation of life-sustaining treatment. Therefore, we could not exclude the risk of self-fulfilling prophecies introducing model prediction bias. 6), The EEG data were collected at different clinical sites, not as part of a single unified study. Therefore, we have evaluated the model performance on the data from independent studies. But the generalization of the proposed model should be further evaluated on more heterogeneous patient cohorts from different clinical centers. 7), The proportion of patients with good outcome was comparable to other studies in the literature [29], [31].

Use of sedatives is common in comatose cardiac arrest patients, however, the effect of sedatives on neurological outcomes have not been quantified, e.g., whether propofol is beneficial or harmful in patients with cell and organ injury after resuscitation from cardiac arrest is unknown. [43] Use of sedatives might have affected the EEG signals used in our prediction models and might impact the generalizability of the study results. [44] Recent studies suggest that the influence of sedatives on EEG patterns does not significantly affect neurological prognostication performance. [37], [45], [46] Nevertheless, usage and dosing of sedative drugs varies across sites, and the effects of propofol and other sedatives in individual critically ill patients varies, thus further investigation of individual-level effects and effect variation across medical centers remains an important topic for investigation.

In the past few decades, neurologic prognostication after cardiac arrest has progressed towards a multimodal paradigm based on integrating information from the clinical examination (e.g. the pupillary light and corneal reflex) with information from other modalities, e.g. somatosensory evoked potential, brain imaging [47]–[50]. Given that different modalities have strengths and weaknesses, multimodality assessments may provide more reliable neurologic prognostication by combining clinical evidence from multiple complementary information sources [7], [51], [52]. Future work on developing more robust multimodal outcome prediction models should focus on well-designed deep learning models that integrate rich, large-scale healthcare data [24], [53], [54] from various institutions to encompass wider practice variations and a broader range of patient phenotypes to improve model performance.

V. Conclusion

In conclusion, we developed a time-sensitive deep learning model for neurological outcome prediction in coma patients after CA with sequences of Bi-LSTMs, which can learn the long-term EEG dynamics during the progressive course of coma recovery. Model performance was evaluated on a large, multicenter, international cohort, and the model showed excellent agreement between its probabilistic predictions and the observed rate of good and poor neurologic outcomes. Our results demonstrate that time-sensitive deep neural networks can extract valuable information from the EEG in patients with coma following cardiac arrest, to provide accurate predictions about the potential recovery of neurologic function.

References

[1] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das *et al.*, “Heart disease and stroke statistics‒2019 update: a report from the american heart association,” *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[2] N. Nielsen, J. Wetterslev, T. Cronberg, D. Erlinge, Y. Gasche, C. Hassager, J. Horn, J. Hovdenes, J. Kjaergaard, M. Kuiper *et al.*, “Targeted temperature management at 33 c versus 36 c after cardiac arrest,” *New England Journal of Medicine*, vol. 369, no. 23, pp. 2197–2206, 2013.

[3] C. W. Callaway, M. W. Donnino, E. L. Fink, R. G. Geocadin, E. Golan, K. B. Kern, M. Leary, W. J. Meurer, M. A. Peberdy, T. M. Thompson *et al.*, “Part 8: post–cardiac arrest care: 2015 american heart association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care,” *Circulation*, vol. 132, no. 18_suppl_2, pp. S465–S482, 2015.

[4] C. Sandroni, A. Cariou, F. Cavallaro, T. Cronberg, H. Friberg, C. Hoedemaekers, J. Horn, J. P. Nolan, A. O. Rossetti, and J. Soar, “Prognostication in comatose survivors of cardiac arrest: an advisory statement from the european resuscitation council and the european society of intensive care medicine,” *Intensive Care Medicine*, vol. 40, no. 12, pp. 1816–1831, 2014.

[5] G. B. Young, “Neurologic prognosis after cardiac arrest,” *New England Journal of Medicine*, vol. 361, no. 6, pp. 605–611, 2009.

[6] D. K. Hahn, R. G. Geocadin, and D. M. Greer, “Quality of evidence in studies evaluating neuroimaging for neurologic prognostication in adult patients resuscitated from cardiac arrest,” *Resuscitation*, vol. 85, no. 2, pp. 165–172, 2014.

[7] A. O. Rossetti, A. A. Rabinstein, and M. Oddo, “Neurological prognostication of outcome in patients in coma after cardiac arrest,” *The Lancet Neurology*, vol. 15, no. 6, pp. 597–609, 2016.

[8] E. F. Wijdicks, A. Hijdra, G. Young, C. Bassetti, and S. Wiebe, “Practice parameter: prediction of outcome in comatose survivors after cardiopulmonary resuscitation (an evidence-based review): report of the quality standards subcommittee of the american academy of neurology,” *Neurology*, vol. 67, no. 2, pp. 203–210, 2006.

[9] B. J. Ruijter, M. C. Tjepkema-Cloostermans, S. C. Tromp, W. M. van den Bergh, N. A. Foudraine, F. H. Kornips, G. Drost, E. Scholten, F. H. Bosch, A. Beishuizen *et al.*, “Early electroencephalography for outcome prediction of postanoxic coma: a prospective cohort study,” *Annals of Neurology*, vol. 86, no. 2, pp. 203–214, 2019.

[10] M. M. Admiraal, A.-F. van Rootselaar, J. Hofmeijer, C. W. Hoedemaekers, C. R. van Kaam, H. M. Keijzer, M. J. van Putten, M. J. Schultz, and J. Horn, “Electroencephalographic reactivity as predictor of neurological outcome in postanoxic coma: a multicenter prospective cohort study,” *Annals of Neurology*, vol. 86, no. 1, pp. 17–27, 2019.

[11] E. Westhall, A. O. Rossetti, A.-F. van Rootselaar, T. W. Kjaer, J. Horn, S. Ullén, H. Friberg, N. Nielsen, I. Rosén, A. Åneman *et al.*, “Standardized EEG interpretation accurately predicts prognosis after cardiac arrest,” *Neurology*, vol. 86, no. 16, pp. 1482–1490, 2016.

[12] J. Hofmeijer, M. C. Tjepkema-Cloostermans, and M. J. van Putten, “Burst-suppression with identical bursts: a distinct EEG pattern with poor outcome in postanoxic coma,” *Clinical Neurophysiology*, vol. 125, no. 5, pp. 947–954, 2014.

[13] M. C. Tjepkema-Cloostermans, J. Hofmeijer, R. J. Trof, M. J. Blans, A. Beishuizen, and M. J. van Putten, “Electroencephalogram predicts outcome in patients with postanoxic coma during mild therapeutic hypothermia,” *Critical Care Medicine*, vol. 43, no. 1, pp. 159–167, 2015.

[14] A. Sivaraju, E. J. Gilmore, C. R. Wira, A. Stevens, N. Rampal, J. J. Moeller, D. M. Greer, L. J. Hirsch, and N. Gaspard, “Prognostication of post-cardiac arrest coma: early clinical and electroencephalographic predictors of outcome,” *Intensive Care Medicine*, vol. 41, no. 7, pp. 1264–1272, 2015.

[15] S. Tsetsou, J. Novy, M. Oddo, and A. O. Rossetti, “EEG reactivity to pain in comatose patients: importance of the stimulus type,” *Resuscitation*, vol. 97, pp. 34–37, 2015.

[16] E. Amorim, J. C. Rittenberger, J. J. Zheng, M. B. Westover, M. E. Baldwin, C. W. Callaway, A. Popescu *et al.*, “Continuous EEG monitoring enhances multimodal outcome prediction in hypoxic–ischemic brain injury,” *Resuscitation*, vol. 109, pp. 121–126, 2016.

[17] J. Jing, A. Herlopian, I. Karakis, M. Ng, J. J. Halford, A. Lam, D. Maus, F. Chan, M. Dolatshahi, C. F. Muniz *et al.*, “Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms,” *JAMA Neurology*, vol. 77, no. 1, pp. 49–57, 2020.

[18] M. C. Hermans, M. B. Westover, M. J. van Putten, L. J. Hirsch, and N. Gaspard, “Quantification of EEG reactivity in comatose patients,” *Clinical Neurophysiology*, vol. 127, no. 1, pp. 571–580, 2016.

[19] N. Gaspard, L. J. Hirsch, S. M. LaRoche, C. D. Hahn, M. B. Westover, and C. C. E. M. R. Consortium, “Interrater agreement for critical care EEG terminology,” *Epilepsia*, vol. 55, no. 9, pp. 1366–1373, 2014.

[20] M. C. Tjepkema-Cloostermans, F. B. van Meulen, G. Meinsma, and M. J. van Putten, “A cerebral recovery index (CRI) for early prognosis in patients after cardiac arrest,” *Critical Care*, vol. 17, no. 5, pp. 1–11, 2013.

[21] M. C. Tjepkema-Cloostermans, J. Hofmeijer, A. Beishuizen, H. W. Hom, M. J. Blans, F. H. Bosch, and M. J. Van Putten, “Cerebral recovery index: reliable help for prediction of neurologic outcome after cardiac arrest,” *Critical Care Medicine*, vol. 45, no. 8, pp. e789–e797, 2017.

[22] S. Lee, X. Zhao, K. A. Davis, A. A. Topjian, B. Litt, and N. S. Abend, “Quantitative EEG predicts outcomes in children after cardiac arrest,” *Neurology*, vol. 92, no. 20, pp. e2329–e2338, 2019.

[23] J. Hofmeijer, T. M. Beernink, F. H. Bosch, A. Beishuizen, M. C. Tjepkema-Cloostermans, and M. J. van Putten, “Early EEG contributes to multimodal outcome prediction of postanoxic coma,” *Neurology*, vol. 85, no. 2, pp. 137–143, 2015.

[24] K.-H. Yu, A. L. Beam, and I. S. Kohane, “Artificial intelligence in healthcare,” *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[25] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[26] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, “Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning,” *Nature Biomedical Engineering*, vol. 2, no. 3, pp. 158–164, 2018.

[27] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed *et al.*, “Do no harm: a roadmap for responsible machine learning for health care,” *Nature Medicine*, vol. 25, no. 9, pp. 1337–1340, 2019.

[28] W.-L. Zheng, H. Sun, O. Akeju, and M. B. Westover, “Adaptive sedation monitoring from EEG in ICU patients with online learning,” *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 6, pp. 1696–1706, 2019.

[29] M. C. Tjepkema-Cloostermans, C. da Silva Lourenço, B. J. Ruijter, S. C. Tromp, G. Drost, F. H. Kornips, A. Beishuizen, F. H. Bosch, J. Hofmeijer, and M. J. van Putten, “Outcome prediction in postanoxic coma with deep learning,” *Critical Care Medicine*, vol. 47, no. 10, pp. 1424–1432, 2019.

[30] E. Amorim, M. Van der Stoel, S. B. Nagaraj, M. M. Ghassemi, J. Jing, U.-M. O'Reilly, B. M. Scirica, J. W. Lee, S. S. Cash, and M. B. Westover, “Quantitative EEG reactivity and machine learning for prognostication in hypoxic-ischemic brain injury,” *Clinical Neurophysiology*, vol. 130, no. 10, pp. 1908–1916, 2019.

[31] M. M. Ghassemi, E. Amorim, T. Alhanai, J. W. Lee, S. T. Herman, A. Sivaraju, N. Gaspard, L. J. Hirsch, B. M. Scirica, S. Biswal *et al.*, “Quantitative electroencephalogram trends predict recovery in hypoxic-ischemic encephalopathy,” *Critical Care Medicine*, vol. 47, no. 10, pp. 1416–1423, 2019.

[32] S. B. Nagaraj, M. C. Tjepkema-Cloostermans, B. J. Ruijter, J. Hofmeijer, and M. J. van Putten, “The revised cerebral recovery index improves predictions of neurological outcome after cardiac arrest,” *Clinical Neurophysiology*, vol. 129, no. 12, pp. 2557–2566, 2018.

[33] C. M. Booth, R. H. Boone, G. Tomlinson, and A. S. Detsky, “Is this patient dead, vegetative, or severely neurologically impaired?: assessing outcome for comatose survivors of cardiac arrest,” *JAMA*, vol. 291, no. 7, pp. 870–879, 2004.

[34] F. S. Taccone, J. Horn, C. Storm, A. Cariou, C. Sandroni, H. Friberg, C. A. Hoedemaekers, and M. Oddo, “Death after awakening from postanoxic coma: the best CPC project,” *Critical Care*, vol. 23, no. 1, pp. 1–8, 2019.

[35] M. B. Westover, M. M. Shafi, S. Ching, J. J. Chemali, P. L. Purdon, S. S. Cash, and E. N. Brown, "Real-time segmentation of burst suppression patterns in critical care EEG monitoring," *Journal of Neuroscience Methods*, vol. 219, no. 1, pp. 131–141, 2013.

[36] J. Jing, H. Sun, J. A. Kim, A. Herlopian, I. Karakis, M. Ng, J. J. Halford, D. Maus, F. Chan, M. Dolatshahi *et al.*, "Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation," *JAMA Neurology*, vol. 77, no. 1, pp. 103–108, 2020.

[37] B. J. Ruijter, J. Hofmeijer, M. C. Tjepkema-Cloostermans, and M. J. van Putten, "The prognostic value of discontinuous EEG patterns in postanoxic coma," *Clinical Neurophysiology*, vol. 129, no. 8, pp. 1534–1543, 2018.

[38] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[39] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[40] C. Cortes and M. Mohri, "Confidence intervals for the area under the ROC curve," *Advances in Neural Information Processing Systems*, vol. 17, pp. 305–312, 2005.

[41] T. Lin, T. Guo, and K. Aberer, "Hybrid neural networks for learning the trend in time series," in *International Joint Conference on Artificial Intelligence*, no. CONF, 2017, pp. 2273–2279.

[42] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, 2019.

[43] R. J. Madathil, R. S. Hira, M. Stoeckl, F. Sterz, J. B. Elrod, and G. Nichol, "Ischemia reperfusion injury as a modifiable therapeutic target for cardioprotection or neuroprotection in patients undergoing cardiopulmonary resuscitation," *Resuscitation*, vol. 105, pp. 85–91, 2016.

[44] E. Amorim, J. C. Rittenberger, M. E. Baldwin, C. W. Callaway, A. Popescu, and P. C. A. Service, "Malignant EEG patterns in cardiac arrest patients treated with targeted temperature management who survive to hospital discharge," *Resuscitation*, vol. 90, pp. 127–132, 2015.

[45] C. M. Drohan, A. I. Cardi, J. C. Rittenberger, A. Popescu, C. W. Callaway, M. E. Baldwin, and J. Elmer, "Effect of sedation on quantitative electroencephalography after cardiac arrest," *Resuscitation*, vol. 124, pp. 132–137, 2018.

[46] B. J. Ruijter, M. J. van Putten, W. M. van den Bergh, S. C. Tromp, and J. Hofmeijer, "Propofol does not affect the reliability of early EEG for outcome prediction of comatose patients after cardiac arrest," *Clinical Neurophysiology*, vol. 130, no. 8, pp. 1263–1270, 2019.

[47] C. S. Youn, C. W. Callaway, J. C. Rittenberger *et al.*, "Combination of initial neurologic examination, quantitative brain imaging and electroencephalography to predict outcome after cardiac arrest," *Resuscitation*, vol. 110, pp. 120–125, 2017.

[48] M. B. Bevers, B. M. Scirica, K. R. Avery, G. V. Henderson, A. P. Lin, and J. W. Lee, "Combination of clinical exam, MRI and EEG to predict outcome following cardiac arrest and targeted temperature management," *Neurocritical Care*, vol. 29, no. 3, pp. 396–403, 2018.

[49] J. H. Kim, M. J. Kim, J. S. You, H. S. Lee, Y. S. Park, I. Park, and S. P. Chung, "Multimodal approach for neurologic prognostication of out-of-hospital cardiac arrest patients undergoing targeted temperature management," *Resuscitation*, vol. 134, pp. 33–40, 2019.

[50] M. Oddo and A. O. Rossetti, "Early multimodal outcome prediction after cardiac arrest in patients treated with hypothermia," *Critical Care Medicine*, vol. 42, no. 6, pp. 1340–1347, 2014.

[51] C. Sandroni, S. DArrigo, and J. P. Nolan, "Prognostication after cardiac arrest," *Critical Care*, vol. 22, no. 1, pp. 1–9, 2018.

[52] N. Ben-Hamouda, F. S. Taccone, A. O. Rossetti, and M. Oddo, "Contemporary approach to neurologic prognostication of coma after cardiac arrest," *Chest*, vol. 146, no. 5, pp. 1375–1386, 2014.

[53] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[54] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.