

Title: Exploration of Short-range Neonatal Seizure Forecasting with Quantitative EEG Based Deep Learning

Authors:

Jonathan Kim BS¹, Hannah C Glass MD MAS^{2,3,4}, Edilberto Amorim MD², Vikram R Rao MD PhD², Danilo Bernardo MD²

¹ School of Medicine, University of Illinois - Urbana Champaign. Chicago, IL

² Department of Neurology, University of California, San Francisco. San Francisco, CA

³ Department of Epidemiology and Biostatistics, University of California, San Francisco. San Francisco, CA

⁴ Department of Pediatrics, University of California, San Francisco. San Francisco, CA

Abstract: 295

Manuscript: 2968

Figures: 6

Supplementary Figures: 3

Supplementary Tables: 3

References: 43

ABSTRACT

Background

In this study, we utilize robust feature selection of quantitative encephalography (QEEG) features for inclusion into a deep learning (DL) model for short-range forecasting of neonatal seizure risk.

Methods

We used publicly available EEG seizure datasets with a total of 132 neonates. The Boruta algorithm with Shapley values was used for QEEG feature selection into a convolutional long short-term memory (ConvLSTM) DL model to classify preictal versus interictal states. ConvLSTM was trained and evaluated with 10-fold cross-validation. Performance was evaluated with varying seizure prediction horizons (SPH) and seizure occurrence periods (SOP).

Results

Boruta with Shapley values identified statistical moments, spectral power distributions, and RQA features as robust predictors of preictal states. ConvLSTM performed best with SPH 3 min and SOP 7 min, demonstrating 80% sensitivity with 36% of time spent in false alarm, AUROC of 0.80, and AUPRC of 0.23. The model demonstrated ECE of 0.106, consistent with moderate calibration. Evaluation of forecasting skill with BSS under varying SPH demonstrated a peak BSS of 0.056 and a trend for decreasing BSS with increasing SPH.

Conclusion

Statistical moments, spectral power, and recurrence quantitative analysis are predictive of the preictal state. Short-range neonatal seizure forecasting is feasible with DL models utilizing these features.

MANUSCRIPT

Introduction

Neonatal seizures, with an incidence rate of one to three per 1000 live births, are associated with substantial long-term morbidity and mortality (1, 2). Prompt seizure treatment is critical for neonates, as a higher seizure burden is associated with increased treatment resistance and mortality (3, 4, 5, 6). A promising strategy to improve neonatal clinical outcomes has focused on identifying seizure-prone neonates using clinical and EEG features to reduce the time to seizure diagnosis and treatment (7, 8, 9, 10, 11).

Recent studies have leveraged machine learning (ML) to predict seizures in neonatal encephalopathy (NE) with high accuracy utilizing long forecast horizons during the acute postnatal period (10, 11). Pavel et al. introduced a neonatal ML model utilizing clinical variables and quantitative EEG (QEEG) features shortly after birth to predict neonates with NE who later developed seizures, forecasting individual seizure risk over several days (10). Recently, McKee et al. developed a ML model on qualitative EEG and clinical features from the first day of life that could predict subsequent seizures during the acute monitoring period spanning days (11). While these and other prior studies have predicted seizure risk over an observation period spanning several days (7, 8, 9, 10, 11), short-range seizure risk forecasting, or forecasts with higher temporal resolution, remains unexplored in the neonatal population.

In contrast to long-range forecasts, short-range forecasting provides more precise and timely information regarding the imminence of seizure onset (Figure 1a). The provision of short-range forecasts may facilitate the investigation of prophylactic interventions in higher seizure risk populations such as NE, and help optimize the allocation of monitoring resources, which in many environments is limited in accessibility to continuous EEG (12, 13, 14). Thus, our objective was to extend upon prior work in neonatal seizure forecasting, by investigating high-temporal-resolution forecasting spanning minutes.

Here, we develop a deep learning (DL) approach for QEEG-based neonatal seizure forecasting and utilize publicly available neonatal EEG datasets to evaluate model performance on short-range forecasting of neonatal seizures. Considering the uncertainty into which specific QEEG features are most predictive of the preictal state in neonates, we integrated robust QEEG feature selection methods into our approach.

Methods

Subject Data:

We utilized two publicly available EEG datasets from Helsinki University Hospital (HUH) and Cork University Maternity Hospital (Cork). The HUH dataset consists of multi-channel 256 Hz EEG recorded from 79 term neonates at the NICU in HUH, Helsinki, Finland, with total 60 hours of recording (15). In the HUH dataset, the presence of seizures in the EEGs was annotated independently by three experts. The most common diagnosis in this dataset was birth asphyxia (35 patients). The Cork dataset from the INFANT Research Center, Cork University Maternity Hospital, contains EEG records from 53 neonates affected by HIE with total 169 hours total

recording (16). While the majority of HUH subjects contained ictal samples, only two Cork subjects have EEG records containing seizures in this dataset, thus providing relatively more balance between seizure-containing and non-seizure containing subjects.

Preprocessing:

We segmented EEG data into 20-second non-overlapping epochs with class labels of preictal, interictal, and ictal states. *Ictal periods* were defined as time periods in which at least 2 experts annotated a seizure. We defined *preictal periods* as between 6 min to 1 min prior to seizure onset and *interictal periods* as between 1 minute after end of seizure to 5 min prior to the next seizure. Right-censored periods in which it is unknown whether a seizure occurred within the prediction window at the end of data recordings were excluded. Prior to windowing and feature calculation, the raw EEG signal was band-pass filtered between 0.1 Hz and 20 Hz, and then resampled from the original 256 Hz to two times the Nyquist frequency, 40 Hz. For seizures that are less than the seizure prediction horizon (SPH) from the previous seizure, we consider them as a single seizure event. To increase model robustness and generalizability, we performed augmentation of the training dataset by translation invariant transforms; specifically, we transposed the raw EEG channels across the frontal-occipital and left-right axes. Candidate QEEG features from feature categories were calculated on non-overlapping 20-second EEG epochs.

Feature Selection:

Feature selection refers to the process of identifying the subset of features most pertinent to a prediction model's performance. Unlike automatic or emergent feature extraction methods, such as those used by convolutional neural networks, QEEG features may vary significantly in relevance and predictive power, necessitating a robust feature selection process. Increasing the number of features without bounds, or excessively high-dimensional QEEG data, may impair model performance due to increased data complexity, a concept broadly known in neuroscience and other domains as the 'curse of dimensionality' (17). Overfitting also becomes a concern, as models trained on a multitude of features may undermine generalizability to novel data. Feature selection, distilling the feature set down to those of true relevance, enhances model generalization, streamlines model training, and improves model interpretation (18).

For feature selection, we utilized BorutaSHAP(19), which integrates the robustness of the Boruta algorithm feature selection strategy with the Shapley value feature importances derived from SHapley Additive exPlanations (20). The Boruta algorithm is a feature selection method used in ML, which is based on the random forest classification algorithm(21). It utilizes feature importances, such as SHAP or Gini, to measure to identify significantly predictive features in a dataset. These feature importances are iteratively compared with those of shadow features, which are generated from random shuffling of the real features to provide a reference. A threshold for feature selection is defined by the maximum importance score derived from the shadow features. Using this threshold, two-sided T-test is used to ascertain the relative significance of each feature—features significantly below the threshold are considered 'unimportant', while those significantly above the threshold are deemed 'important'. This feature importance ranking utilizes Shapley values, a game-theoretic method that determines individual feature contributions to model predictions, which provides consistent, accurate feature importance scores. Shapley values represent each feature's average marginal contribution to model prediction, across all possible combinations(20). Through this integrated BorutaSHAP procedure, we identified the top feature

categories consistently surpassing the threshold, thereby indicating significant correlation with the preictal or interictal state.

Selected QEEG features from the top three feature categories are demonstrated in Supplementary Table 1 in more detail, which included standard summary statistics (mean, standard deviation, kurtosis, skew, the 10th percentile, and the 90th percentile), calculated across each montaged channel left-right pair, and power spectral features calculated at each montage channel (22). Other feature categories previously utilized for seizure prediction in prior studies that were evaluated included autocorrelation, entropy, detrended fluctuation analysis, and coherence (23).

Model Design:

We developed a custom convolutional long short-term memory neural network (ConvLSTM), an architecture that has previously been utilized for seizure prediction (24). An advantage of utilizing an LSTM-based architecture relative to conventional ML methods is their capability to learn underlying temporal dependencies from sequential data. The incorporation of the convolutional layer allows for local temporal feature extraction. Further details regarding ConvLSTM architecture are shown in Supplementary Figure 1.

Model Training:

As described above, the data used to train the model consisted of three classes, preictal, interictal, and ictal. The loss function used to train the model was formulated as a multi-class cross-entropy function that disregarded performance on the ictal class while penalizing mislabeling those members of preictal and interictal classes. The pytorch and pytorch-lightning libraries were used to evaluate models(25). Grid-search optimization was used to tune ConvLSTM and demonstrated best-performance with a learning rate of $5e-5$, a maximum of 100 training epochs, and early stopping of training conditioned on validation loss delta of $1e-6$. For each fold of the 10-fold cross-validation, the entire dataset was split into one train/validation set consisting of 90% of the data and one test set consisting of 10% of the data. The train/validation set was then further split into a training set and a validation set, with the training set taking 75% of the train/validation set and the validation set taking the remaining 25%. All splits were created with inter-subject stratification, ensuring that data from a given subject was either solely in the train set or solely in the test set.

Model Performance Evaluation:

We employed the seizure alarm framework, which utilizes the seizure prediction horizon (SPH) and seizure occurrence period (SOP) (26). In this framework, a minimum SPH is designated to ensure sufficient lead time preceding a seizure to allow for timely intervention strategies to be employed. Concurrently, a maximum SOP is set to the desired temporal resolution suited to the duration of applicability of a given forecast. If the predicted risk is above the designated threshold, an alarm is triggered, lasting the combined duration of SPH and SOP. Details regarding the definitions of SPH and SOP are shown in Supplementary Figure 2. We evaluated area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), Expected Calibration Error (ECE), and Brier-Skill Score (BSS). PR-AUC illustrates the trade-off between precision (the proportion of true positives among all predicted positives) and recall (the proportion of true positives among all actual positives). It is advantageous in imbalanced datasets such as in this case, where the number of interictal periods (negative class) significantly exceeds

preictal periods (positive class), as it focuses on the model's performance concerning the positive class. ECE is evaluated to measure the reliability of the model's predicted probabilities and is commonly used to evaluate neural network performance (27). It assesses the discrepancy between the predicted and true probabilities of the outcomes, which is important in determining whether the model is well-calibrated. The BSS assesses the model's forecasting skill relative to a reference random classifier model. The BSS considers both calibration and discrimination of the model. A positive BSS indicates that the model performs better than the reference model, while a negative BSS signifies the opposite. For the BSS calculation, we utilized the standard climatology reference, which accounts for the prevalence of the positive class (preictal states). For comparison of ConvLSTM to conventional ML methods, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and random forest (RF) classifiers, we additionally evaluated F1 score and Matthew Correlation Coefficient (MCC).

RESULTS

Study Subjects

The HUH EEG dataset consists of 79 subjects with median age 40 weeks (interquartile range: 39.4 - 40.7) with 39 (49%) subjects with total of 516 across all subjects (28). The Cork EEG dataset consists of 53 subjects with median age 39.5 (37.8-40.5) weeks with two subjects (4%) who had seizures (16). The average duration per subjects of the HUH and Cork datasets were 1.2 (1.1-1.6) hrs and 3 (2-4) hours.

The study overview is demonstrated in Figure 1b, including all feature categories considered for feature selection. Feature selection with Boruta algorithm utilizing Shapley value feature importances demonstrated that the top three feature categories included statistical moments, spectral power, and RQA features (Figure 2). These features were identified as consistently demonstrating importance scores exceeding the threshold (Supplementary Figure 3), indicating their predictive relevance. QEEG features from the top three feature categories (Supplementary Table 1) were incorporated into ConvLSTM. Examples of the resulting time-varying seizure risk forecasts for individual neonates with seizures are shown in Figures 3. In these examples, periods containing heightened seizure risk occur with varying lead times prior to the seizure occurrences (red lines). In comparison, the selected examples from neonates with no seizures show no peaks indicative of preictal state (Figure 4).

To assess the performance of ConvLSTM over varying short-range forecast horizons, we evaluated area under the ROC curve and area under the AUPRC varying SOP and SPH between 1 to 7 minutes (Figure 5). Peak AUROC was 0.80 with SPH of 3 minutes and SOP of 7 minutes, and at these SPH and SOP settings, with ROC threshold adjusted to correspond with 80% sensitivity, there was a corresponding 36% time spent during false alarm. The peak AUPRC was 0.23 with SPH of 1 minute and SOP of 7 minutes. For any given SPH, increasing the SOP led to improved AUROC and AUPRC, reflective of increased ease of forecasting at lower temporal resolutions.

The calibration of the ConvLSTM model was evaluated using the reliability plot and Expected Calibration Error (ECE) metrics. The reliability plot, (Figure 6a), demonstrates the relationship

between the predicted probabilities and the observed frequency of the preictal class. There is a clustering of points below the diagonal, which signifies that ConvLSTM was occasionally overconfident in its predictions for the preictal class at these probabilities. Additionally, isolated points distant from the diagonal occurred at the lowest and highest probability bins, consistent with respective isolated underconfident and overconfident forecasts at these bins. Concordant to the reliability plot findings, the ECE value of 0.106 is consistent with a moderately well-calibrated model. The value of 0.106 is slightly above the general desirable range for a strongly calibrated model, considered less than 0.1 (zero indicates perfect calibration, and one is the maximum value indicative of weak calibration).

To evaluate the forecasting skill of ConvLSTM, we examined the impact of forecasting horizon and resolution utilizing the Brier Skill Score (BSS), which measures the difference between the accuracy of the model's predictions and the accuracy of a reference forecast. The effect of varying SPH and SOP on BSS is shown in Figure 6b. The highest BSS was 0.056 obtained with a SPH of 1 minute and SOP of 10 minutes. We find that for all SOP, as the SPH increases, then model performance concomitantly decreases. This follows the intuition that forecasting farther into the future is inherently more difficult. Similarly, as the SOP temporal resolution becomes finer, model performance also decreases, suggesting that forecasting with increased temporal resolution is also more difficult.

Lastly, we compared ConvLSTM performance to conventional ML methods. ConvLSTM model achieved higher AUROC, AUPRC, and F1 scores than SVM, KNN, logistic regression, and random forest classifiers at the classification of preictal versus interictal states (Supplementary Table 2).

DISCUSSION

In this study, we utilized the Boruta algorithm and Shapley values for robust selection of QEEG features, identifying statistical moments, spectral power, and RQA features as most predictive of interictal and preictal states. We incorporated these features into ConvLSTM and demonstrated the accuracy of providing short-term forecasts of neonatal seizures. Forecasting horizons as brief as 5 minutes has been reported in adult patients with long-term chronic intracranial EEG recordings (23, 29). However, to our knowledge, this is the first study to demonstrate short-range forecasting in the neonatal population. Short-range forecasting in the neonates at high risk of impending seizure may enable time-sensitive interventions in higher seizure risk populations, such as in neonatal encephalopathy, and help optimize the allocation of monitoring resources.

Notably, recent ML approaches at predicting neonatal seizure risk have utilized a combination of QEEG features with clinical features with ensemble ML methods such as gradient boosted decision trees and random forests to estimate seizure risk accurately (10, 11). In contrast to these prior works, which estimate neonatal seizure risk over several days, our approach focuses on short-term forecasting with a time resolution of minutes.

Recent seizure forecasting studies in the pediatric population have utilized subject-specific feature engineering and model training for their analyses (30, 31, 32, 33). Notably, Tsiouras et al.

achieved 100% sensitivity with an FPR of 0.06/hr on the full CHB-MIT set by using a subject-specific neural network with a long short-term memory (LSTM) architecture (34). Our approach is distinct from these prior works in that we developed a subject-independent model instead of a subject-specific model. Although our performance metrics are moderately lower than those attained in prior patient-specific models, the subject-independent approach presented obviates the need for subject-specific training. This is particularly clinically advantageous for neonates who are often at immediate elevated risk of seizure after birth, such as in cases of neonatal encephalopathy, where subject-specific training data may not be immediately available before the first seizure. Furthermore, subject-independent models facilitate efficient resource allocation, as it can be readily implemented at different sites without fine-tuning, reducing the need for specialized expertise and computational resources.

Concerning our modeling approach, we utilized a convolutional LSTM that demonstrated overall improved performance compared to conventional ML methods. ConvLSTM incorporates strengths of both convolutional neural network (CNN) and LSTM architectures: the convolutional module can efficiently downsample the input signal while extracting local temporal features predictive of seizure risk, whereas the LSTM module effectively learns long-range temporal dependencies. Prior studies in seizure prediction have previously utilized CNN (33, 35, 36), LSTM (34, 37), and CNN-LSTM(24). In contrast to the previously published CNN-LSTM method, which utilized CNN-LSTM on short-time Fourier-transformed (STFT) EEG signal, we incorporated other QEEG features predictive of preictal states in addition to spectral power changes, including statistical moments and RQA features.

Regarding the limitations of our study, we recognize that we utilized a relatively low number of subjects and did not utilize a held-out or independent evaluation dataset. Thus, validation on larger and independent datasets are necessary to confirm our findings. Additionally, our model demonstrated a relatively higher false alarm rate and lower sensitivity than prior studies, which have predominantly utilized the CHB-MIT dataset (Supplementary Table 3), which may be attributed to our subject-independent approach. In addition, the relatively lower performance seen in this study may reflect the usage of shorter-range SOP and SPH, as forecasting with increasingly higher temporal resolution is considered more difficult in complex systems, such as in seismology or meteorology, and this has also been suggested for seizure forecasting (38, 39, 40).

In conclusion, we demonstrate the potential of applying ML approaches to enable time-dependent neonatal seizure forecasting, facilitating more precise timing and temporal understanding of neonatal seizure susceptibility.

REFERENCES

1. Glass HC, Grinspan ZM, Shellhaas RA. Outcomes after acute symptomatic seizures in neonates. *Semin Fetal Neonatal Med.* 2018;23(3):218-22.
2. Ronen GM, Buckley D, Penney S, Streiner DL. Long-term prognosis in children with neonatal seizures: a population-based study. *Neurology.* 2007;69(19):1816-22.
3. Glass HC, Shellhaas RA, Wusthoff CJ, Chang T, Abend NS, Chu CJ, et al. Contemporary Profile of Seizures in Neonates: A Prospective Cohort Study. *J Pediatr.* 2016;174:98-103 e1.
4. Glass HC, Hong KJ, Rogers EE, Jeremy RJ, Bonifacio SL, Sullivan JE, et al. Risk factors for epilepsy in children with neonatal encephalopathy. *Pediatr Res.* 2011;70(5):535-40.
5. Glass HC, Soul JS, Chu CJ, Massey SL, Wusthoff CJ, Chang T, et al. Response to antiseizure medications in neonates with acute symptomatic seizures. *Epilepsia.* 2019;60(3):e20-e4.
6. Painter MJ, Scher MS, Stein AD, Armatti S, Wang Z, Gardiner JC, et al. Phenobarbital compared with phenytoin for the treatment of neonatal seizures. *The New England Journal of Medicine.* 1999;341(7):485-9.
7. Glass HC, Wusthoff CJ, Shellhaas RA, Tsuchida TN, Bonifacio SL, Cordeiro M, et al. Risk factors for EEG seizures in neonates treated with hypothermia: a multicenter cohort study. *Neurology.* 2014;82(14):1239-44.
8. Sansevere AJ, Kapur K, Peters JM, Fernandez IS, Loddenkemper T, Soul JS. Seizure Prediction Models in the Neonatal Intensive Care Unit. *J Clin Neurophysiol.* 2019;36(3):186-94.
9. Cornet M-C, Pasupuleti A, Fang A, Gonzalez F, Shimotake T, Ferriero DM, et al. Predictive value of early EEG for seizures in neonates with hypoxic-ischemic encephalopathy undergoing therapeutic hypothermia. *Pediatric Research.* 2018;84(3):399-402.
10. Pavel AM, O'Toole JM, Proietti J, Livingstone V, Mitra S, Marnane WP, et al. Machine learning for the early prediction of infants with electrographic seizures in neonatal hypoxic-ischaemic encephalopathy. *Epilepsia.* 2022.
11. McKee JL, Kaufman MC, Gonzalez AK, Fitzgerald MP, Massey SL, Fung F, et al. Leveraging electronic medical record-embedded standardised electroencephalogram reporting to develop neonatal seizure prediction models: a retrospective cohort study. *The Lancet Digital Health.* 2023;5(4):e217-e26.
12. Mathieson SR, Stevenson NJ, Low E, Marnane WP, Rennie JM, Temko A, et al. Validation of an automated seizure detection algorithm for term neonates. *Clin Neurophysiol.* 2016;127(1):156-68.
13. Rennie J, Boylan G. Treatment of neonatal seizures. *Archives of Disease in Childhood - Fetal and Neonatal Edition.* 2007;92(2):F148-F50.
14. Hochberg O, Berger I. Bedside EEG Monitoring in the Neonatal Intensive Care Unit. *Current Treatment Options in Pediatrics.* 2022;8(3):295-307.
15. Stevenson NJ, Tapani K, Lauronen L, Vanhatalo S. A dataset of neonatal EEG recordings with seizure annotations. *Sci Data.* 2019;6:190039.
16. O'toole JM, Mathieson SR, Raurale SA, Magarelli F, Marnane WP, Lightbody G, et al. Neonatal EEG graded for severity of background abnormalities in hypoxic-ischaemic encephalopathy. *Scientific Data.* 2023;10(1):129.
17. Allen EA, Erhardt EB, Calhoun VD. Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron.* 2012;74(4):603-8.

18. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: A data perspective. *ACM computing surveys (CSUR)*. 2017;50(6):1-45.
19. Keany E. BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. <https://doi.org/105281/zenodo4247618>. 2020;1.1.
20. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30.
21. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *Journal of statistical software*. 2010;36:1-13.
22. Gao Y, Chen X, Liu A, Liang D, Wu L, Qian R, et al. Pediatric seizure prediction in scalp EEG using a multi-scale neural network with dilated convolutions. *IEEE journal of translational engineering in health and medicine*. 2022;10:1-9.
23. Kuhlmann L, Karoly P, Freestone DR, Brinkmann BH, Temko A, Barachant A, et al. Epilepsyecosystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG. *Brain*. 2018;141(9):2619-30.
24. Shahbazi M, Aghajan H, editors. A generalizable model for seizure prediction based on deep learning using CNN-LSTM architecture. 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP); 2018: IEEE.
25. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*. 2019;32.
26. Maiwald T, Winterhalder M, Aschenbrenner-Scheibe R, Voss HU, Schulze-Bonhage A, Timmer J. Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic. *Physica D: nonlinear phenomena*. 2004;194(3-4):357-68.
27. Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D, editors. Measuring Calibration in Deep Learning. *CVPR workshops*; 2019.
28. Stevenson NJ, Tapani K, Lauronen L, Vanhatalo S. A dataset of neonatal EEG recordings with seizure annotations. *Scientific Data*. 2019;6(1):190039.
29. Cook MJ, O'Brien TJ, Berkovic SF, Murphy M, Morokoff A, Fabinyi G, et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *The Lancet Neurology*. 2013;12(6):563-71.
30. Zhang Z, Parhi KK. Low-complexity seizure prediction from iEEG/sEEG using spectral power and ratios of spectral power. *IEEE transactions on biomedical circuits and systems*. 2015;10(3):693-706.
31. Daoud H, Bayoumi MA. Efficient epileptic seizure prediction based on deep learning. *IEEE transactions on biomedical circuits and systems*. 2019;13(5):804-13.
32. Alotaiby TN, Alshebeili SA, Alotaibi FM, Alrshoud SR. Epileptic seizure prediction using CSP and LDA for scalp EEG signals. *Computational intelligence and neuroscience*. 2017;2017.
33. Truong ND, Nguyen AD, Kuhlmann L, Bonyadi MR, Yang J, Ippolito S, et al. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Networks*. 2018;105:104-11.
34. Tsiouris KM, Pezoulas VC, Zervakis M, Konitsiotis S, Koutsouris DD, Fotiadis DI. A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. *Computers in biology and medicine*. 2018;99:24-37.

35. Khan H, Marcuse L, Fields M, Swann K, Yener B. Focal onset seizure prediction using convolutional networks. *IEEE Transactions on Biomedical Engineering*. 2017;65(9):2109-18.
36. Zhang Y, Guo Y, Yang P, Chen W, Lo B. Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network. *IEEE Journal of Biomedical and Health Informatics*. 2019;24(2):465-74.
37. Ali H, Karim F, Qureshi JJ, Abuassba AO, Bulbul MF, editors. Seizure prediction using bidirectional LSTM. *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health: International 2019 Cyberspace Congress, CyberDI and CyberLife, Beijing, China, December 16–18, 2019, Proceedings, Part I 3*; 2019: Springer.
38. Bauer P, Thorpe A, Brunet G. The quiet revolution of numerical weather prediction. *Nature*. 2015;525(7567):47-55.
39. Beroza GC, Segou M, Mostafa Mousavi S. Machine learning and earthquake forecasting—next steps. *Nat Commun*. 2021;12(1):4761.
40. Baud MO, Proix T, Gregg NM, Brinkmann BH, Nurse ES, Cook MJ, et al. Seizure forecasting: bifurcations in the long and winding road. *Epilepsia*. 2022.
41. Golding B. Nimrod: A system for generating automated very short range forecasts. *Meteorological Applications*. 1998;5(1):1-16.
42. Rawald T, Sips M, Marwan N. PyRQA—Conducting recurrence quantification analysis on very long time series efficiently. *Computers & Geosciences*. 2017;104:101-8.

FIGURES

Figure 1

- a. The left panel illustrates differences between short- and long-range forecasting. Short-range forecasts may be updated frequently (indicated by dynamics in the dashed line), which enables these forecasts to be revised regularly as new data becomes available. Current neonatal long-range seizure forecasting typically provide time-invariant, or static forecasts. The right panel illustrates the relation between forecast information content and forecasting lead time, following analogous forecasting definitions previously developed in meteorology (41). Here, *information content* may pertain to the degree of accuracy, certainty, and overall applicability of the forecasts. There is a downward trend in information content in the theoretical limit of predictability, short- and long-range forecast curves reflecting that in a complex and chaotic systems such as the brain, there is an inevitable increase in uncertainty as one attempts to forecasts further into the future with increasing lead time, or SPH. The short-range forecast may potentially contain higher information content than long-range forecasts during short lead times because it is based on more up to date observations. However, as the SPH becomes longer, for both short- and long-range forecasts, the information content decreases due to increased uncertainty and a reduction in the accuracy of forecasts. The long-range forecast curve starts after one hour after birth reflecting current long-range forecasting models that are based on information obtained from at least the first hour of EEG information.
- b. Starting from a broad selection of features across different categories, feature selection utilizing the Boruta algorithm utilizing Shapley value feature importances was performed to identify features most predictive of interictal and preictal states. Feature categories in which had

significant importance were incorporated into ConvLSTM to estimate time-varying, short-range seizure risk.

Feature Abbreviations: Recurrence quantification analysis (RQA), Detrended Fluctuation Analysis (DFA), Autoregressive Modeling (AR), Statistical Moments (Stat Moments)

Metric Abbreviations: Area Under Receiver Operator Characteristic Curve (AUROC), Area Under Precision Recall Curve (AUPRC)

Figure 2

Boruta analysis was performed to identify the QEEG feature categories most predictive at classifying interictal and preictal states. The top feature categories included features from statistical moments (e.g. Standard deviation (STD) asymmetric index (AI)), spectral power distributions in different frequency bands, and RQA feature categories such as RQA recurrence rate. Features are ranked in accordance by their Shapley feature importance scores, which were utilized in the Boruta analysis. Abbreviations and formulations for features from the top 3 feature categories are defined in Supplementary Table 1.

Figure 3

Examples from 3 subjects with seizures are shown. Prior to seizure occurrences, there are increases in estimated preictal state probability. There are also elevations in preictal state probability not associated with immediate seizure.

Figure 4

Examples from 3 subjects without seizures are shown. There are no significant elevations in preictal state probability seen.

Figure 5

Area under the ROC and PR curve analyses are demonstrated for ConvLSTM model. A) shows respective AUROC's for SOP and SPH varied between 1 and 7 minutes. The best AUROC was 0.80 at SOP of 7 minutes and SPH of 3 minutes. B) shows respective AUPRC's for SOP and SPH varied between 1 and 7 minutes. The best AUPRC was 0.23 at SOP of 7 minutes and SPH of 3 minutes.

Figure 6

6a demonstrates ConvLSTM calibration evaluation using a reliability plot and Expected Calibration Error (ECE) metric. The ECE plot An ECE (ϵ) value of 0.106 indicates a moderately-well calibrated model.

6b demonstrates the influence of varying SPH and SOP on forecasting skill as evaluated by the Brier Skill Score (BSS). The best BSS occurred with SPH of 1 minute and SOP of 10 minutes. An increase in SPH correlates with a decline in BSS, and analogously, as the SOP decreases, model performance also decreases, indicating that forecasting with increasing lead-times and increasing temporal resolution becomes more challenging.

Supplementary Figure 1

The ConvLSTM architecture consisted of a convolutional layer with 256 output filters, followed by a batch normalization layer, then followed by an LSTM module with 256 cells and one

hidden layer with size of 32, followed by a rectified linear unit (ReLU), followed by a 50% dropout layer, a fully connected linear layer (FCN), and then a final ReLU layer as final output. A softmax function was applied to the final ReLU layer to yield probabilistic predictions for each output class. For visualization purposes only 64 filters/cells are shown as opposed to 256.

Supplementary Figure 2

The seizure prediction horizon (SPH) and seizure occurrence period (SOP) evaluation framework considers that there should be a minimum SPH to provide ample lead time before a seizure to allow for intervention and that the alarm should have SOP selected to align forecast duration with the specified clinical observation period. The system triggers an alarm, lasting the combined duration of SPH and SOP, if the designated seizure threshold is met. At time t , a true positive alarm occurs if a seizure initiates between $t+\text{SPH}$ and $t+\text{SPH}+\text{SOP}$; otherwise, a false positive is marked. A false negative occurs if a seizure occurs at time t_s and no alarm is activated. A true negative occurs when no alarm is triggered and no seizure occurs.

Supplementary Figure 3

Boruta analysis was performed to identify the most predictive QEEG features at each channel. In the top 20 features the mean Shapley values were predominantly from statistical moments, spectral power distribution, and RQA feature categories. Abbreviations for these top 3 feature categories are defined in Supplementary Table 1.

SUPPLEMENTARY TABLES

Supplementary Table 1

| Family | Abbreviation | Description |
|------------------------------------|--------------|---|
| Recurrence Quantification Analysis | RQA | We utilized features derived from Recurrence Quantification Analysis (RQA), a nonlinear data analysis technique based upon the recurrence plot, which is a plot of the recurrence of states or patterns in a time series (42). In each 20-second window, a recurrence plot was calculated for each montaged channel. The RQA features below (e.g. recurrence rate, determinism, laminarity, trapping time) were calculated on each per-channel recurrence plot. |
| | RR | Recurrence rate |
| | DET | Determinism |
| | LAM | Laminarity |
| | L_max | Longest diagonal line |
| | L_entr | Entropy of diagonal lines |
| | L_mean | Average diagonal line |
| | TT | Trapping time |
| Asymmetry Indices | | For each corresponding pair of channels, L and R, mirrored across the vertical axis, we calculate an asymmetry index utilizing the following formula: $\frac{ fn(L) - fn(R) }{fn(L) + fn(R)}$; 'fn' denotes a set of functions below (fn_mean, fn_std, fn_kurt, etc...) In this formula, 'fn' signifies a set of functions — (fn_mean) mean, standard deviation (fn_std), kurtosis (fn_kurt), skewness (fn_skew), the tenth percentile (fn_ten), and the |

| | | |
|----------------|-------------|---|
| | | ninetieth percentile (fn_ninety) — each of which is applied independently to the values of channels L and R. We then calculate the average of these function outcomes for each corresponding pair of channels. The goal is to quantify the differences between these mirrored (L/R) EEG channels. |
| | Asymm_mean | Mean (fn_mean) |
| | Asymm_std | Standard deviation (fn_std) |
| | Asymm_kurt | Kurtosis (fn_kurt) |
| | Asymm_skew | Skewness (fn_skew) |
| | Asymm_p10 | 10th percentile (fn_ten) |
| | Asymm_p90 | 90th percentile (fn_ninety) |
| Spectral power | Power_delta | Relative spectral power between 0.1 and 4 Hz |
| | Power_theta | Relative spectral power between 4 and 8 Hz |
| | Power_alpha | Relative spectral power between 8 and 12 Hz |
| | Power_beta | Relative spectral power between 12 and 40 Hz |

Supplementary Table 2

| | AUROC | AUPRC | MCC | F1 |
|------------------------|---------------------|---------------------|---------------------|---------------------|
| ConvLSTM | 0.678 ±0.041 | 0.218 ±0.059 | 0.255 ±0.054 | 0.334 ±0.050 |
| Random Forest | 0.612 ±0.021 | 0.147 ±0.023 | 0.141 ±0.024 | 0.252 ±0.030 |
| Support Vector Machine | 0.613 ±0.037 | 0.163 ±0.030 | 0.144 ±0.030 | 0.262 ±0.031 |
| Logistic Regression | 0.598 ±0.025 | 0.163 ±0.027 | 0.112 ±0.023 | 0.243 ±0.027 |
| K-Nearest Neighbors | 0.633 ±0.022 | 0.112 ±0.013 | 0.151 ±0.023 | 0.255 ±0.029 |

Data reported as average performance across all cross-validation folds (10) with ± standard error mean. Abbreviations: Matthew Correlation Coefficient (MCC). F1 score (F1).

Supplementary Table 3

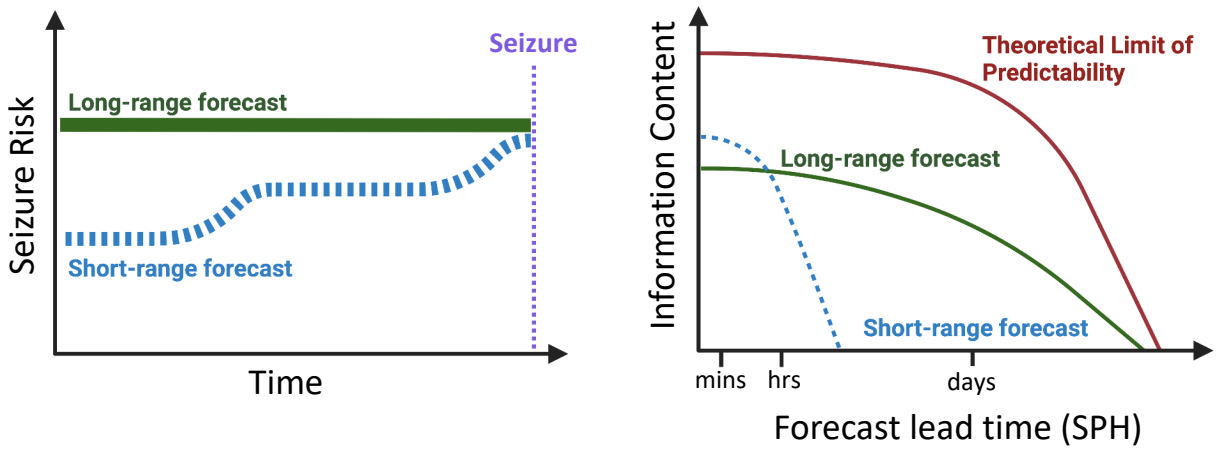
| Year | Authors | Dataset | Feats | CLF | Same Cal | # Sz | Sens (%) | FPR (/h) | SOP (min) | SPH (min) |
|------|-----------------|------------------|--|------|----------|------|----------|----------|-----------|-----------|
| 2016 | Zhang & Parhi | MIT, 17 patients | power spectral density ratio | SVM | no | 80 | 98.68 | 0.05 | 50 | 0** |
| 2017 | Alotaiby et al | MIT, 23 patients | CSP | LDA | yes | 170 | 81 | 0.47 | 60 | 0 |
| 2018 | Khan et al | MIT, 15 patients | wavelet transform | CNN | yes | 18 | 83.33 | 0.15 | 10 | 0** |
| 2018 | Truong et al | MIT, 13 patients | short-time Fourier transform | CNN | yes | 64 | 81.2 | 0.16 | 30 | 5 |
| 2018 | Tsiouras et al | MIT, 23 patients | time domain, frequency domain, graph theory features, correlation features | LSTM | no | 185 | 100 | 0.06 | 30 | 0** |
| 2019 | Daoud & Bayoumi | MIT, 8 patients | DCAE + Bi-LSTM | | yes | 43 | 99.6 | 0.004 | 60 | 0** |

**SPH implicitly set to 0 in these works

Abbreviations: Features (Feats), Classifier (CLF), Seizures (Sz), Sensitivity (Sens), Same Feature engineering (FE), Same Calibration (Cal) denotes studies which were calibrated on the same dataset used for training.

Figure 1

a Short-range vs long-range neonatal seizure forecasts



b

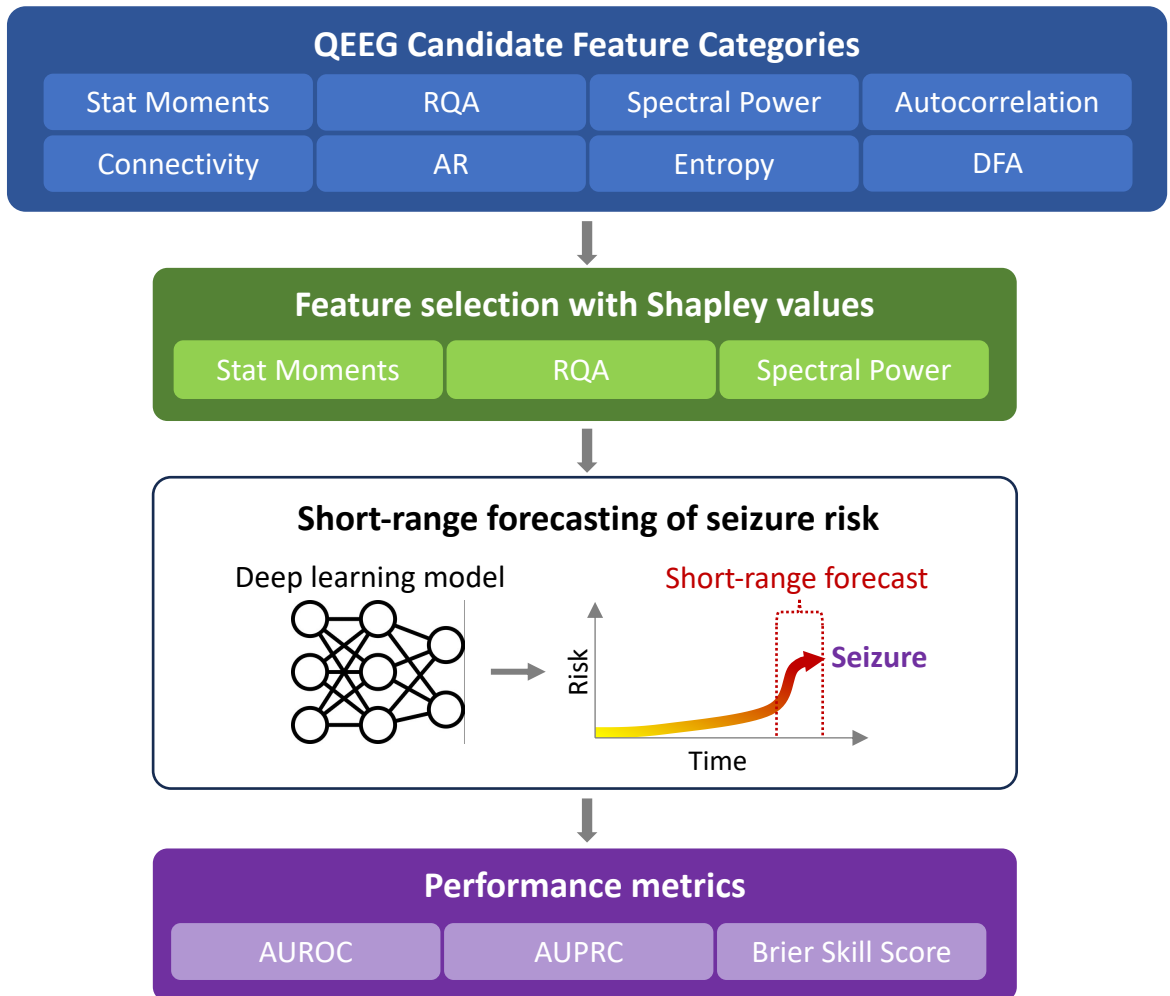


Figure 2

Shapley Feature Importance by Category (Top 10)

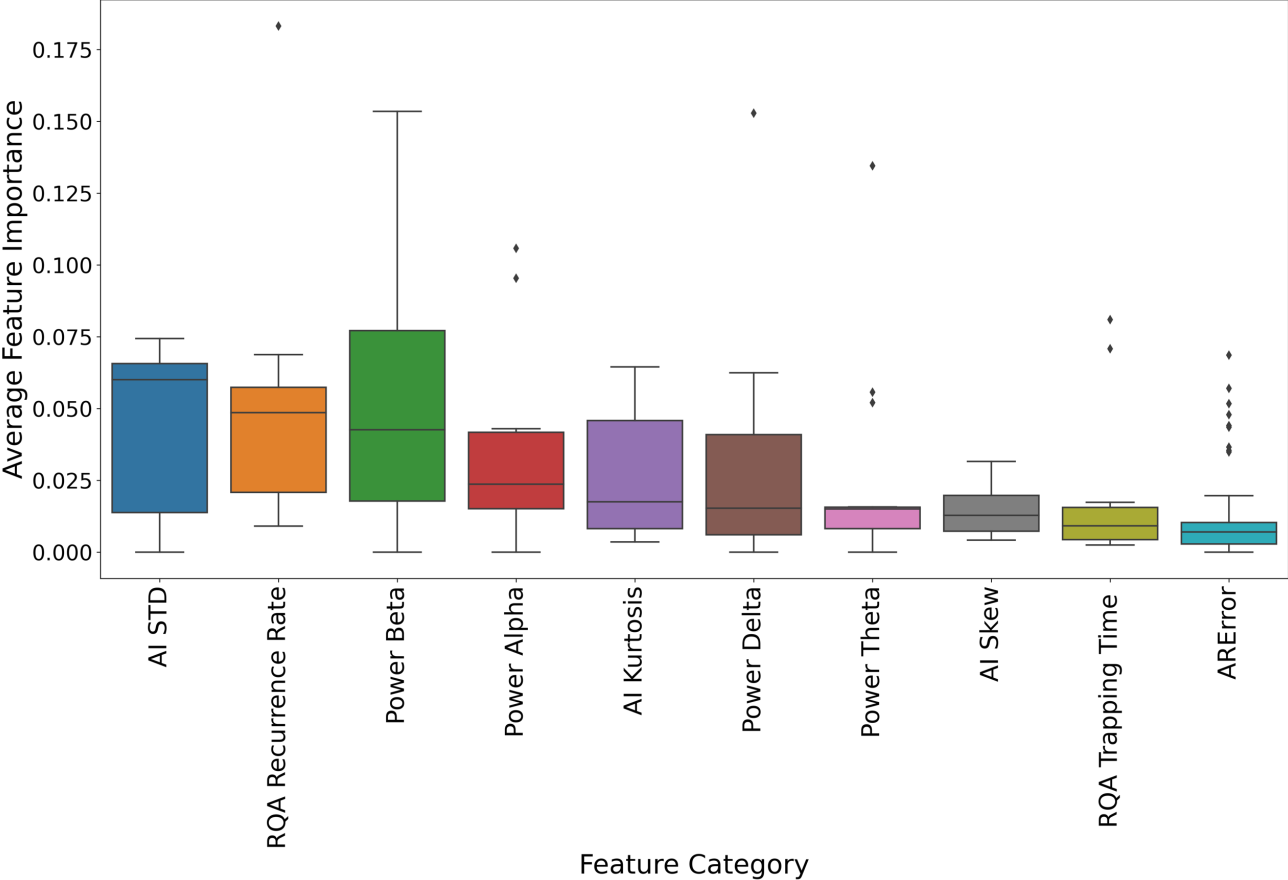


Figure 3

Subjects with seizures

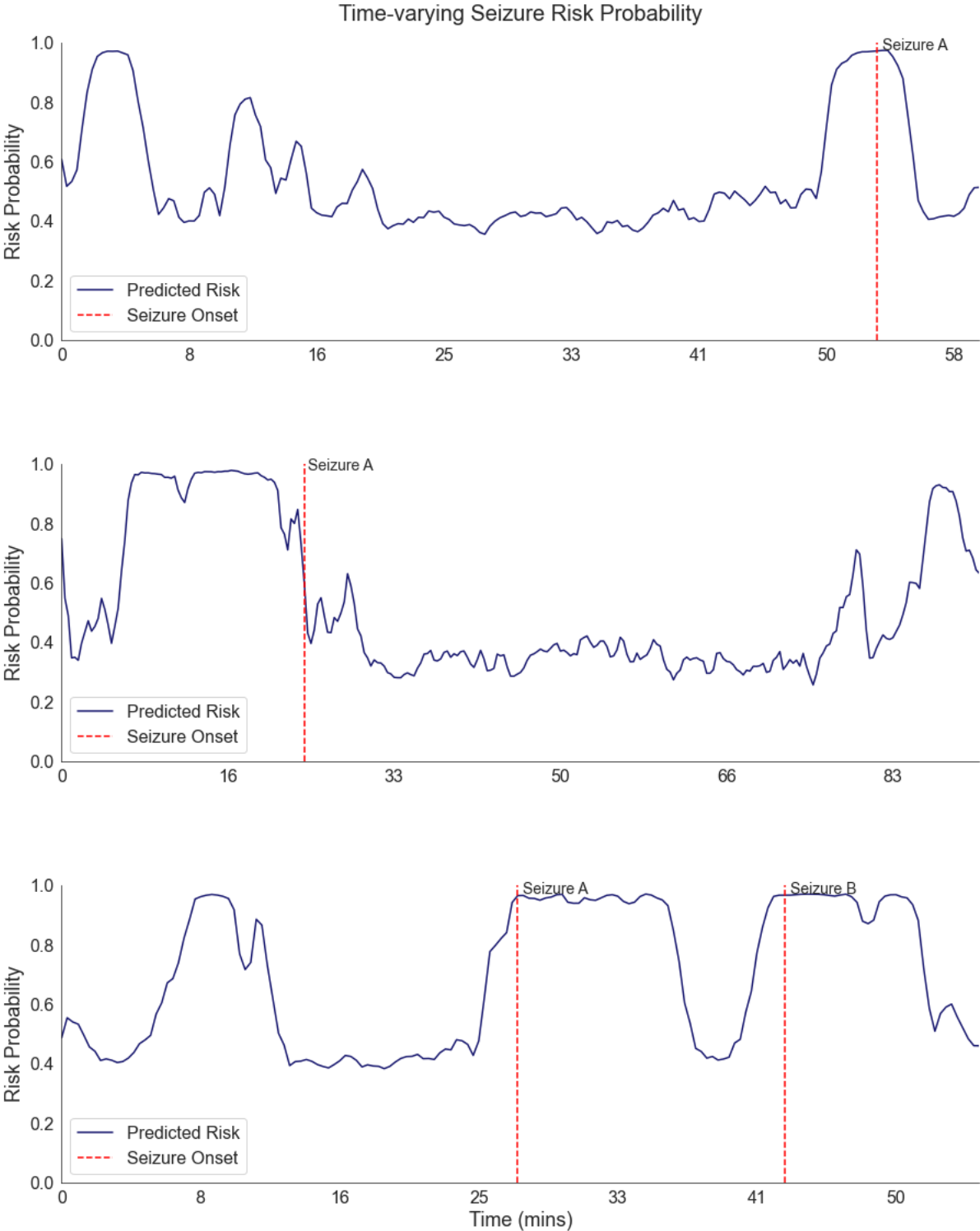


Figure 4

Subjects without seizures

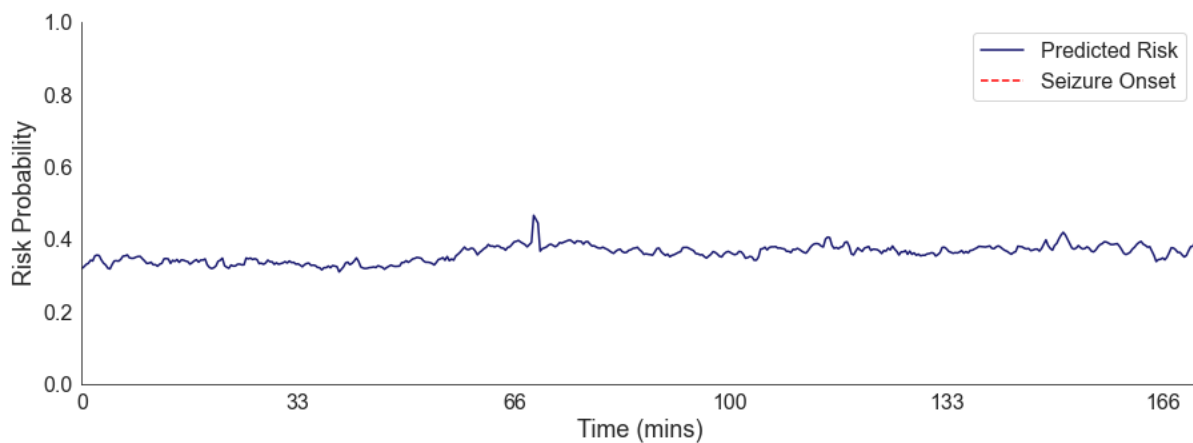
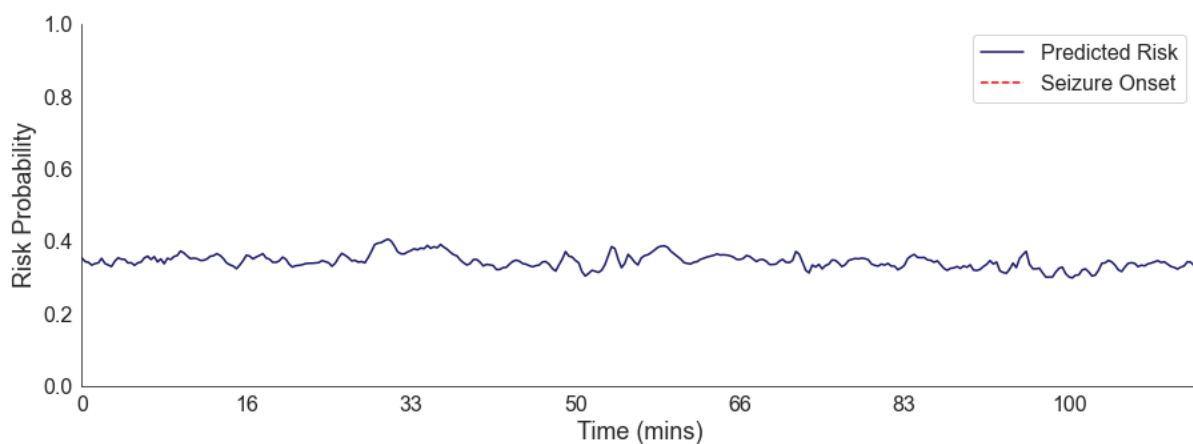
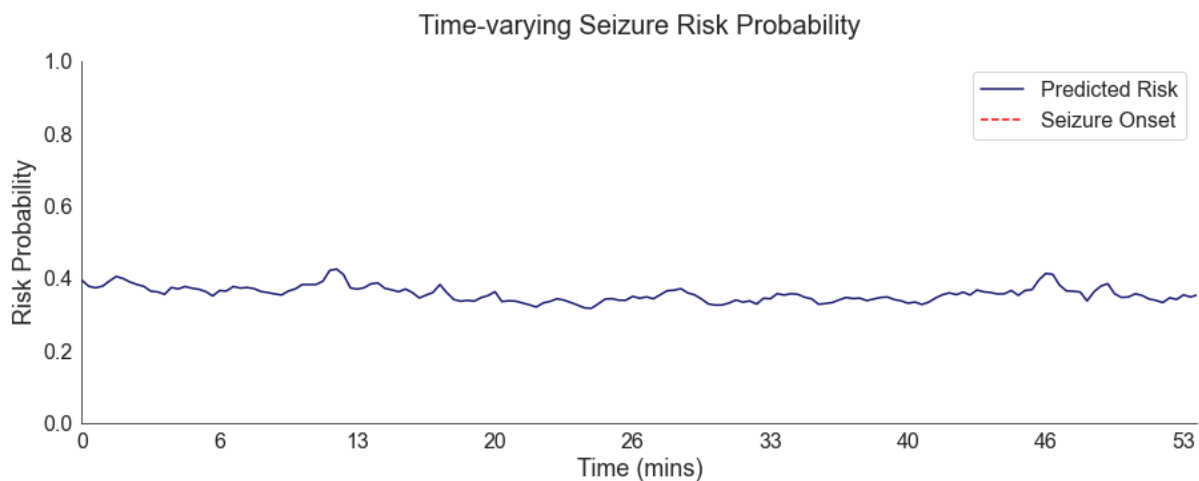
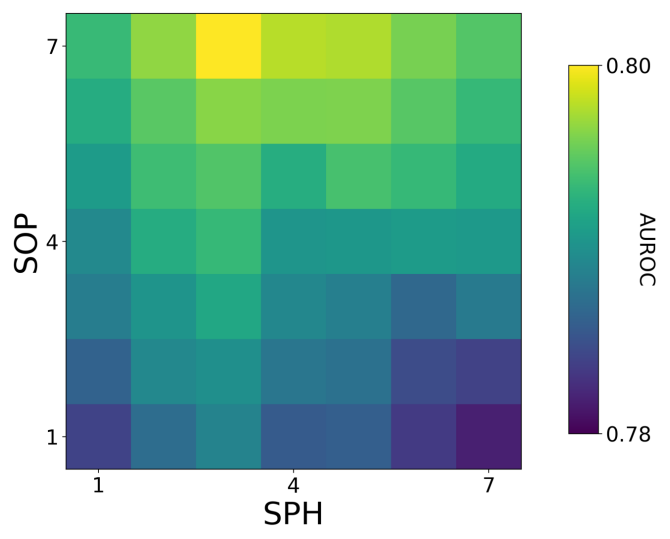


Figure 5

a AUROC Scores under varying SOP and SPH



b AUPRC Scores under varying SOP and SPH

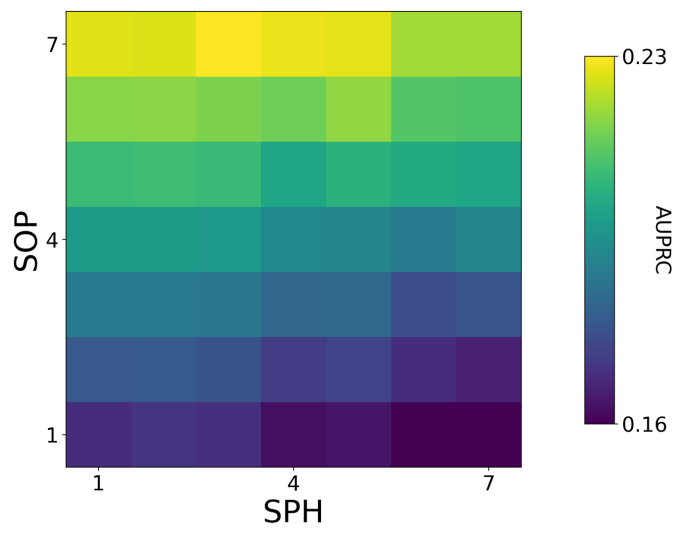
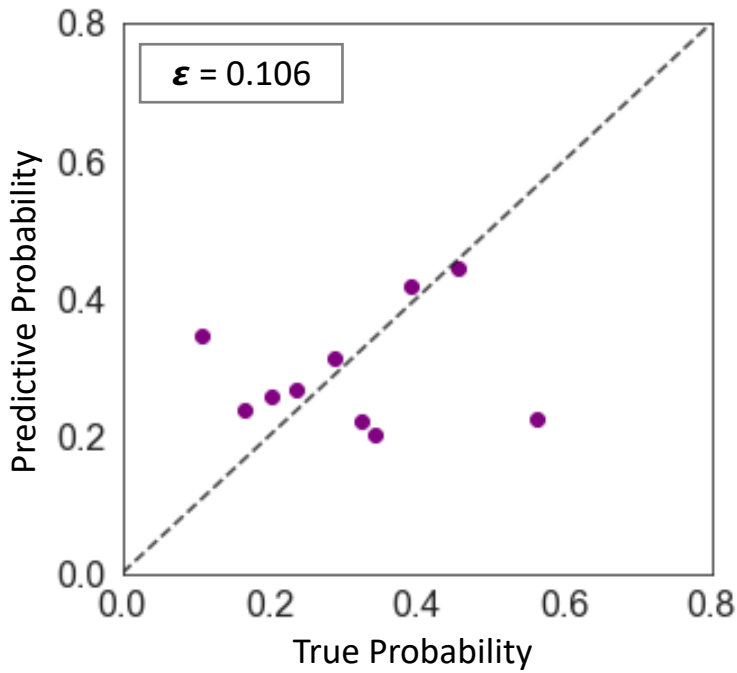
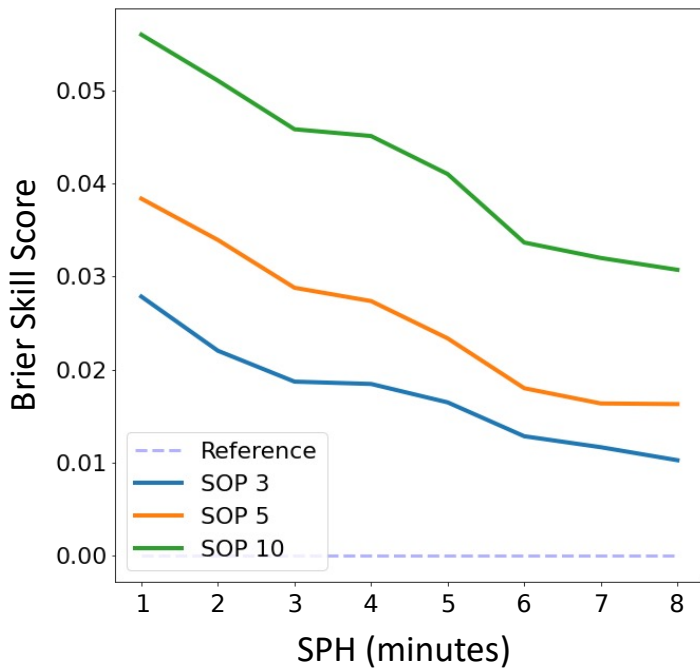


Figure 6

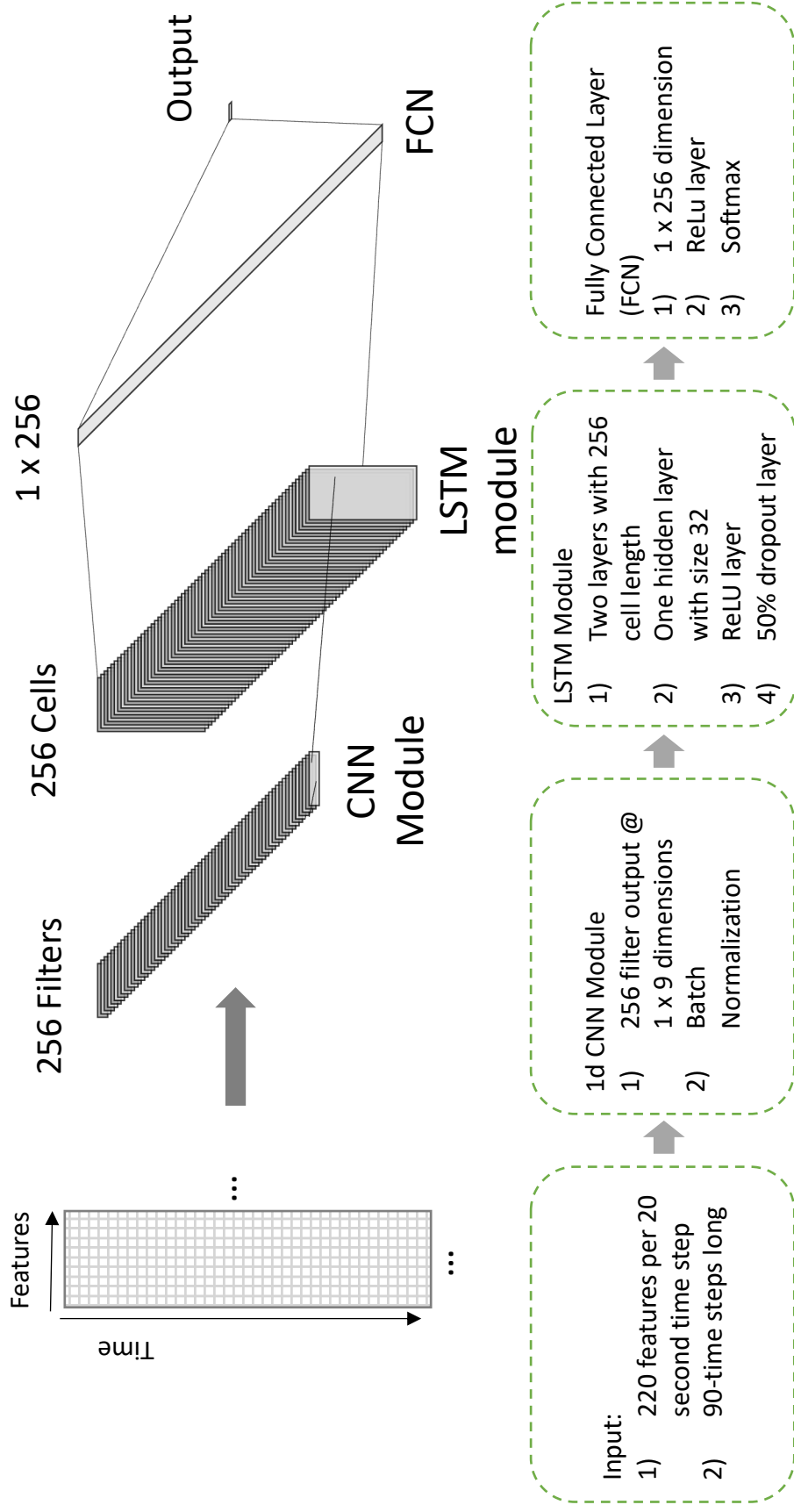
a Reliability Plot and Expected Calibration Error (ϵ)



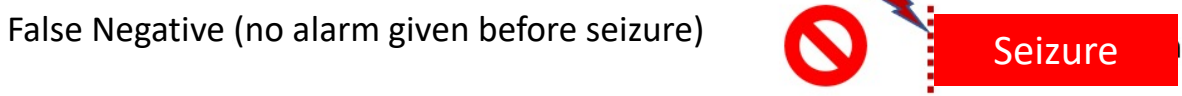
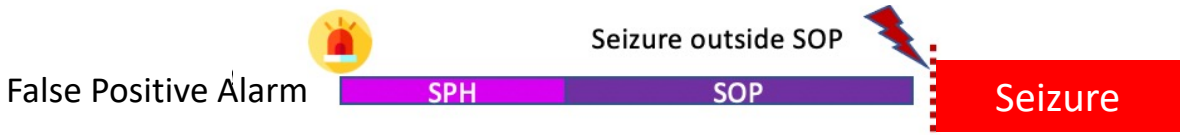
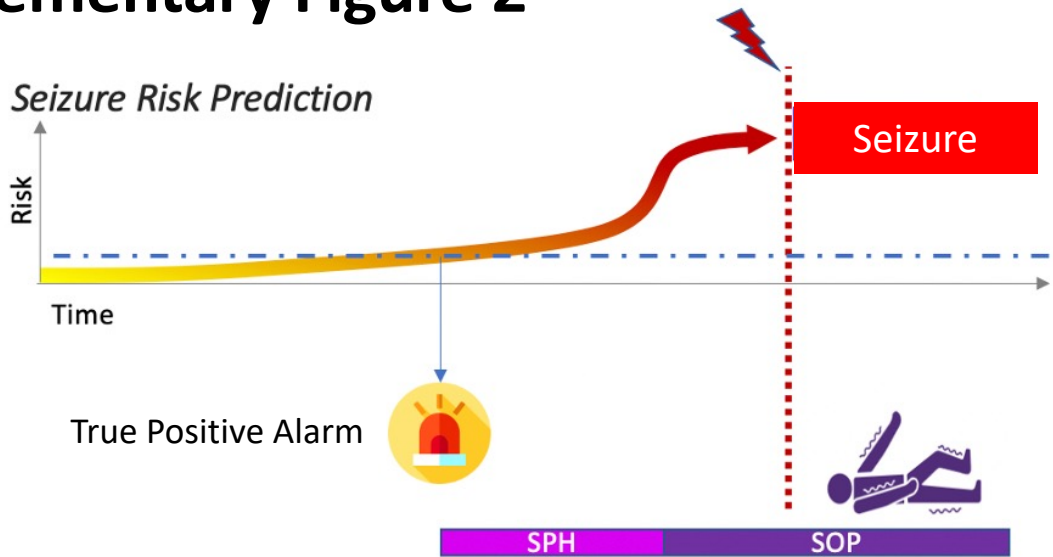
b Brier Skill Score dependence on SPH and SOP



Supplementary Figure 1: ConvLSTM Architecture



Supplementary Figure 2



Supplementary Figure 3

Shapley Feature Importance (top 20 of 400 shown)

