NEJM AI

NEJM AI 2024; 1 (6)
DOI: 10.1056/AIoa2300331

ORIGINAL ARTICLE

# Improving Clinician Performance in Classifying EEG Patterns on the Ictal–Interictal Injury Continuum Using Interpretable Machine Learning

Alina Jade Barnett, Ph.D.,[1] Zhicheng Guo,[2] Jin Jing, Ph.D.,[3] Wendong Ge, Ph.D.,[3] Peter W. Kaplan, M.B.B.S.,[4] Wan Yee Kong, M.D.,[3] Ioannis Karakis, M.D., Ph.D.,[5] Aline Herlopian, M.D.,[6] Lakshman Arcot Jayagopal, M.D.,[7] Olga Taraschenko, M.D., Ph.D.,[7] Olga Selioutski, D.O.,[8] Gamaleldin Osman, M.D.,[9] Daniel Goldenholz, M.D., Ph.D.,[3] Cynthia Rudin, Ph.D.,[10] and M. Brandon Westover, M.D., Ph.D.[3]

Received: December 22, 2023; Revised: February 12, 2024; Accepted: March 27, 2024; Published: May 23, 2024

## Abstract

**BACKGROUND** In intensive care units (ICUs), critically ill patients are monitored with electroencephalography (EEG) to prevent serious brain injury. EEG monitoring is constrained by clinician availability, and EEG interpretation can be subjective and prone to interobserver variability. Automated deep-learning systems for EEG could reduce human bias and accelerate the diagnostic process. However, existing uninterpretable (black-box) deep-learning models are untrustworthy, difficult to troubleshoot, and lack accountability in real-world applications, leading to a lack of both trust and adoption by clinicians.

**METHODS** We developed an interpretable deep-learning system that accurately classifies six patterns of potentially harmful EEG activity — seizure, lateralized periodic discharges (LPDs), generalized periodic discharges (GPDs), lateralized rhythmic delta activity (LRDA), generalized rhythmic delta activity (GRDA), and other patterns — while providing faithful case-based explanations of its predictions. The model was trained on 50,697 total 50-second continuous EEG samples collected from 2711 patients in the ICU between July 2006 and March 2020 at Massachusetts General Hospital. EEG samples were labeled as one of the six EEG patterns by 124 domain experts and trained annotators. To evaluate the model, we asked eight medical professionals with relevant backgrounds to classify 100 EEG samples into the six pattern categories — once with and once without artificial intelligence (AI) assistance — and we assessed the assistive power of this interpretable system by comparing the diagnostic accuracy of the two methods. The model's discriminatory performance was evaluated with area under the receiver-operating characteristic curve (AUROC) and area under the precision–recall curve. The model's interpretability was measured with task-specific neighborhood agreement statistics that interrogated the similarities of samples and features. In a separate analysis, the latent space of the neural network was visualized by

*Dr. Barnett, Mr. Guo, and Dr. Jing and Drs. Rudin and Westover contributed equally to this article.*

*The author affiliations are listed at the end of the article.*

*Mr. Guo can be contacted at zhicheng.guo@duke.edu or at Duke University Pratt School of Engineering, Box 90291, Durham, NC 27708.*

using dimension reduction techniques to examine whether the ictal–interictal injury continuum hypothesis, which asserts that seizures and seizure-like patterns of brain activity lie along a spectrum, is supported by data.

**RESULTS** The performance of all users significantly improved when provided with AI assistance. Mean user diagnostic accuracy improved from 47 to 71% (P<0.04). The model achieved AUROCs of 0.87, 0.93, 0.96, 0.92, 0.93, and 0.80 for the classes seizure, LPD, GPD, LRDA, GRDA, and other patterns, respectively. This performance was significantly higher than that of a corresponding uninterpretable black-box model (with P<0.0001). Videos traversing the ictal–interictal injury manifold from dimension reduction (a two-dimensional representation of the original high-dimensional feature space) give insight into the layout of EEG patterns within the network's latent space and illuminate relationships between EEG patterns that were previously hypothesized but had not yet been shown explicitly. These results indicate that the ictal–interictal injury continuum hypothesis is supported by data.

**CONCLUSIONS** Users showed significant pattern classification accuracy improvement with the assistance of this interpretable deep-learning model. The interpretable design facilitates effective human–AI collaboration; this system may improve diagnosis and patient care in clinical settings. The model may also provide a better understanding of how EEG patterns relate to each other along the ictal–interictal injury continuum. (Funded by the National Science Foundation, National Institutes of Health, and others.)

## Introduction

Seizure or status epilepticus is found in 20% of patients with severe medical and neurologic illness who undergo brain monitoring with electroencephalography (EEG) because of altered mental status,[1,2] and every hour of seizures detected on EEG further increases the risk of permanent disability or death.[3,4] Even more common, intermediate seizure-like patterns of brain activity, consisting of periodic discharges or rhythmic activity, occur in nearly 40% of patients undergoing EEG monitoring.[5] Two recent studies found evidence that, similar to seizures, this type of activity also increases the risk of disability and death if it persists for a prolonged period.[6,7]

The ictal–interictal injury continuum (IIIC) hypothesis,[8] which posits that these ambiguous brain-wave activities and seizures lie along a spectrum, provides a conceptual framework for understanding these potentially harmful EEG patterns, but categorization in clinical settings remains a challenge. Until recently, manual review of the EEG has been the only method to quantify IIIC EEG activities and patterns, an approach that suffers from subjectivity due to the ambiguous nature of these patterns.[8,9] (In this article, the terms "seizures and seizure-like events" and "IIIC EEG patterns" are used interchangeably.)

Recently, progress in deep learning and the availability of large EEG datasets have made possible the development of automated algorithms to detect and classify seizures[10-13] and other EEG patterns,[14] with one recent model achieving a level of accuracy comparable to that of physician experts.[15] However, a lack of interpretability in many of the previous models' decision-making processes renders them unsuitable for assisting human practitioners with medical decision-making at the point of care. (In this article, "interpretability" means that the model can explain predictions in a way that humans can understand.) Uninterpretable or black-box models, which cannot provide an explanation of their decision-making, are prone to silent failures during clinical operations due to either poor generalization or overreliance on trivial medically irrelevant features.[16,17]

These failures can lead to misdiagnoses and increased risks for patients. As a result, the U.S. Food and Drug Administration and the European Union (through the General Data Protection Regulation) have published new requirements and guidelines calling for interpretability and explainability in artificial intelligence (AI) used for medical applications.[18-20] Although explainability techniques such as Gradient-Weighted Class Activation Mapping and Shapley Additive Explanations,[21-25] try to elucidate model decisions post hoc — meaning that the model architecture, development, and training are completed before applying methods to explain the model[26] — these methods only approximate model reasoning. Consequently, different methods often give conflicting explanations, even when used on the same model and sample. This approach contrasts with a model that has built-in interpretability, whereby the prediction is a direct result of the reasoning; that is, the explanation exactly matches the predictor network's underlying calculations.

Our objective was to build an AI assistive tool for IIIC EEG pattern classification to reduce human subjectivity and

improve user accuracy in practice with an interpretability-focused approach. We aimed to better assist clinicians in classifying EEG patterns accurately and reliably, which are the crucial first steps in the EEG reading process.[27] We also hoped to gain insight into the relationships among EEG patterns and develop evidence related to the IIIC hypothesis.

We introduce a novel interpretable deep-learning algorithm to classify seizures and rhythmic and periodic EEG patterns. We propose an explanation method named "This EEG Looks Like That EEG," abbreviated as TEEGLLTEEG. Our proposed interpretable algorithm outperforms the current state-of-the art black-box IIIC EEG pattern classification algorithm in both classification performance and interpretability metrics. To the best of our knowledge, this is the first body of work to develop an inherently interpretable model for EEG signals. We show the clinical utility of our model in a retrospective analysis, wherein all eight users significantly improved in pattern categorization when provided with AI assistance relative to without AI. In addition, we mapped the network's latent space into two dimensions using a dimension-reduction algorithm, revealing that EEG patterns within the IIIC — despite being given distinct class names — do not exist in isolated islands. Rather, each class is connected to every other class through a sequence of transitional intermediate patterns, which we show in a series of videos. These results lend support to the IIIC hypothesis, which asserts that seizures and seizure-like patterns of brain activity lie along a spectrum.

# Methods

## EEG DATA AND EXPERT LABELS

Our network, called ProtoPMed-EEG, was trained and tested on a large-scale EEG study[28] consisting of 50,697 events from 2711 patients hospitalized between July 2006 and March 2020 who underwent continuous EEG as part of clinical care at Massachusetts General Hospital. The large group was intended to ensure broad coverage of all variations of IIIC events encountered in practice. A total of 124 EEG raters from 18 centers labeled the middle 10 seconds of 50-second EEG segments. Raters produced one of six labels: seizure, lateralized periodic discharges (LPDs), generalized periodic discharges (GPDs), lateralized rhythmic delta activity (LRDA), generalized rhythmic delta activity (GRDA), and other patterns. Other included all patterns (including normal) except seizures and the four rhythmic and periodic patterns (LPD, GPD, LRDA, and GRDA). Patterns obscured by artifacts were scored by experts in the same way that it is done in clinical practice so that when artifacts were present but experts were still able to discern that one of the five target patterns was present, they were instructed to assign the target pattern as the label. The data-labeling procedure is described further in the Supplementary Appendix (Section H).

Mean rater-to-rater interrater reliability (IRR) was moderate (agreement, 52%; kappa, 42%), and mean rater-to-majority IRR was substantial (agreement, 65%; kappa, 61%). Because expert annotators do not always agree, we evaluated our model against the majority vote, whereby the class selected by the majority of raters is the ground truth for each sample. The dataset was split into approximately equally sized training and test sets by patient identification to avoid leakage. Rather than allowing any training set sample to become a prototype, we limited our prototype candidates to 10,641 samples that were thoroughly examined in the data-labeling process ($\geq$20 expert votes).

## INTERPRETABILITY THROUGH MODEL DESIGN

An overview of our model architecture design is shown in the upper panel of Figure 1. The model learned the feature extractor (initialized with weights from Jing et al.[14]), the prototype layer, and the final linear layer. In this work, the prototypes were divided into two categories: single-class prototypes and dual-class prototypes. Single-class prototypes represent EEG patterns that can be clearly attributed to one of the six classes described earlier. However, as described in the IIIC hypothesis, some EEG signals may exist in an intermediate state (e.g., between LPD and seizure or between GRDA and GPD). Therefore, dual-class prototypes represent signals in such intermediate states. Dual-class prototypes are novel to this work. Each learned prototype corresponds to an actual EEG sample from the prototype subset. At test time, the latent feature of each input signal was compared against that of the learned prototypes by calculating their angular distances. The distances were passed through the last linear layer to produce prediction scores (logits), which indicated the model's confidence in its predictions for the EEG class. A more detailed version of the training process and the model architecture is provided in the Supplementary Appendix (Section A).

We show three modes of explanation provided by the model design in the lower panel of Figure 1: latent space explanations (Fig. 1A), decision space explanations (Fig. 1B), and

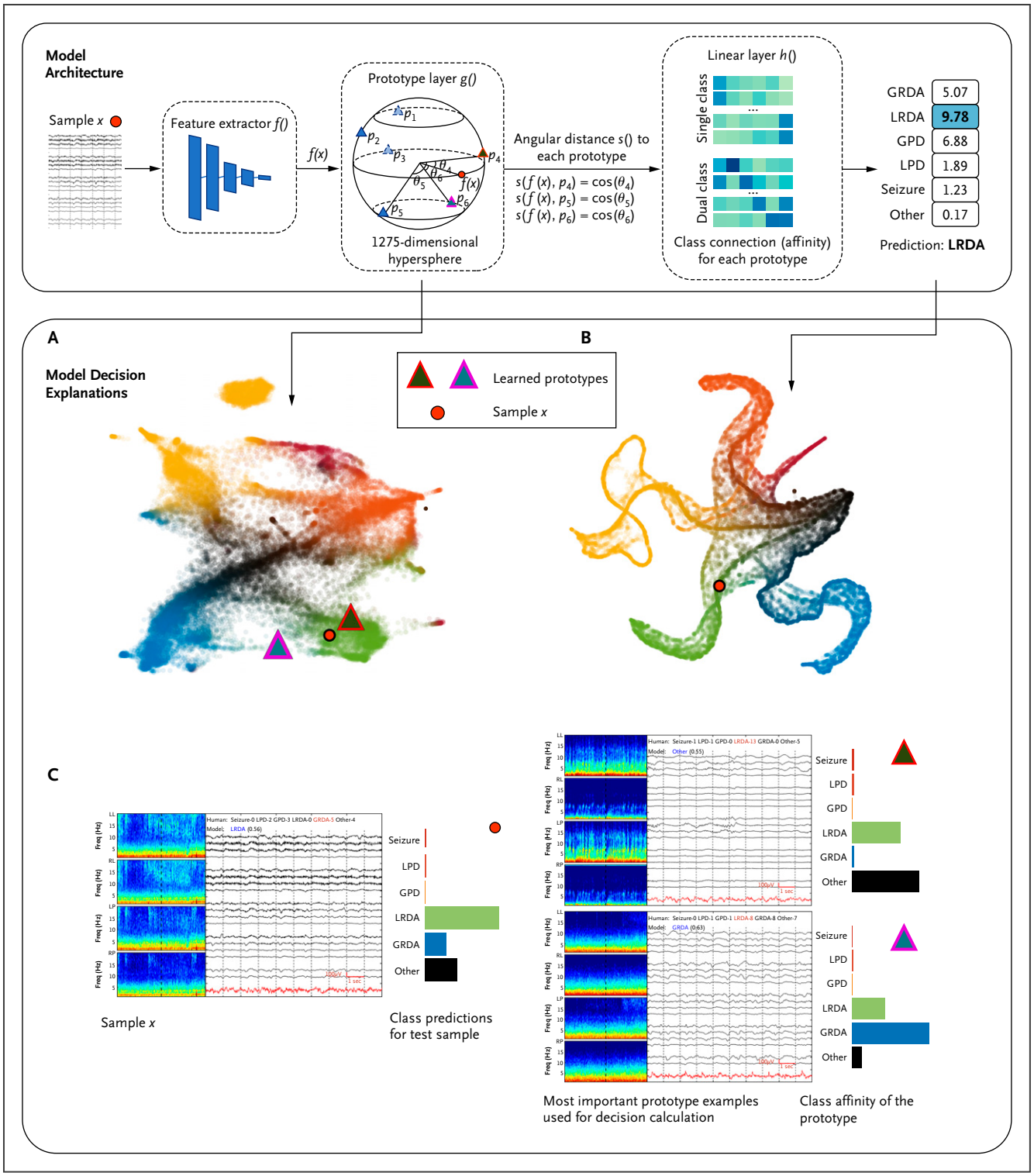


Figure 1. Model Architecture and Decision Explanations.

The input sample *x* is passed through a feature extractor *f* () and a prototype layer *g*() (upper panel; as in Chen et al.[29]). The prototype layer calculates angular distances (as in Donnelly et al.[30]) between the sample feature and the prototypes. The angular distances are multiplied with class affinity to generate the logits (class scores). The softmax calculation converts the logits into prediction probabilities. In the lower panels, three different ways an end user can see how the model reasons about the test sample are shown: latent space explanations (Panel A), decision space explanations (Panel B), and scoring system explanations (Panel C). Freq denotes frequency; GPD, generalized periodic discharge; GRDA, generalized rhythmic delta activity; LL, left lateral; LP, left parasagittal; LPD, lateralized periodic discharge; LRDA, lateralized rhythmic delta activity; RL, right lateral; and RP, right parasagittal.

scoring system explanations (Fig. 1C). (We define interpretable models as models that "explain their predictions in a way that humans can understand."[26]) Every prediction made by our model follows the same logic as the explanation provided by the model. This means that the model explanations have perfect fidelity to the underlying decision-making process by design.

Figure 1A shows how the model perceives the test sample relative to previous cases by projecting the 1275-dimensional latent features to a human-comprehensible two-dimensional space. Figure 1B shows the model's final classification of the test sample relative to the classifications of previous cases. Figure 1C shows how the model uses case-based reasoning to make its prediction (i.e., using previous examples to reason about a new case). This result is achieved by learning a set of prototype samples that are representative of each single-class or dual-class category. Specifically, the model measures the similarity between a new case and the learned prototypical samples. Each explanation is of the form "this sample is class X because it is similar to these prototypes of class X and not similar to prototypes of other classes." The three modes of explanation are integrated into the final graphical user interface (GUI). A snapshot of the dedicated GUI is shown in Figure 2. More details about the GUI are provided in the Supplementary Appendix (Section C).

## USER STUDY

The potential clinical value of the proposed model was assessed with a multiuser study. The study cohort comprised clinical practitioners, including a nurse; an EEG technician; and medical doctors pursuing or having completed residency or fellowship training in neurology, stroke, dermatology, and neurophysiology. None of the participants had expertise in machine learning, and none were EEG experts (i.e., none were physicians who had completed clinical neurophysiology fellowship training). As such, they were an ideal cohort to represent a real-world user population as the interpretable system aims to assist clinical practitioners without prerequisites for AI knowledge.

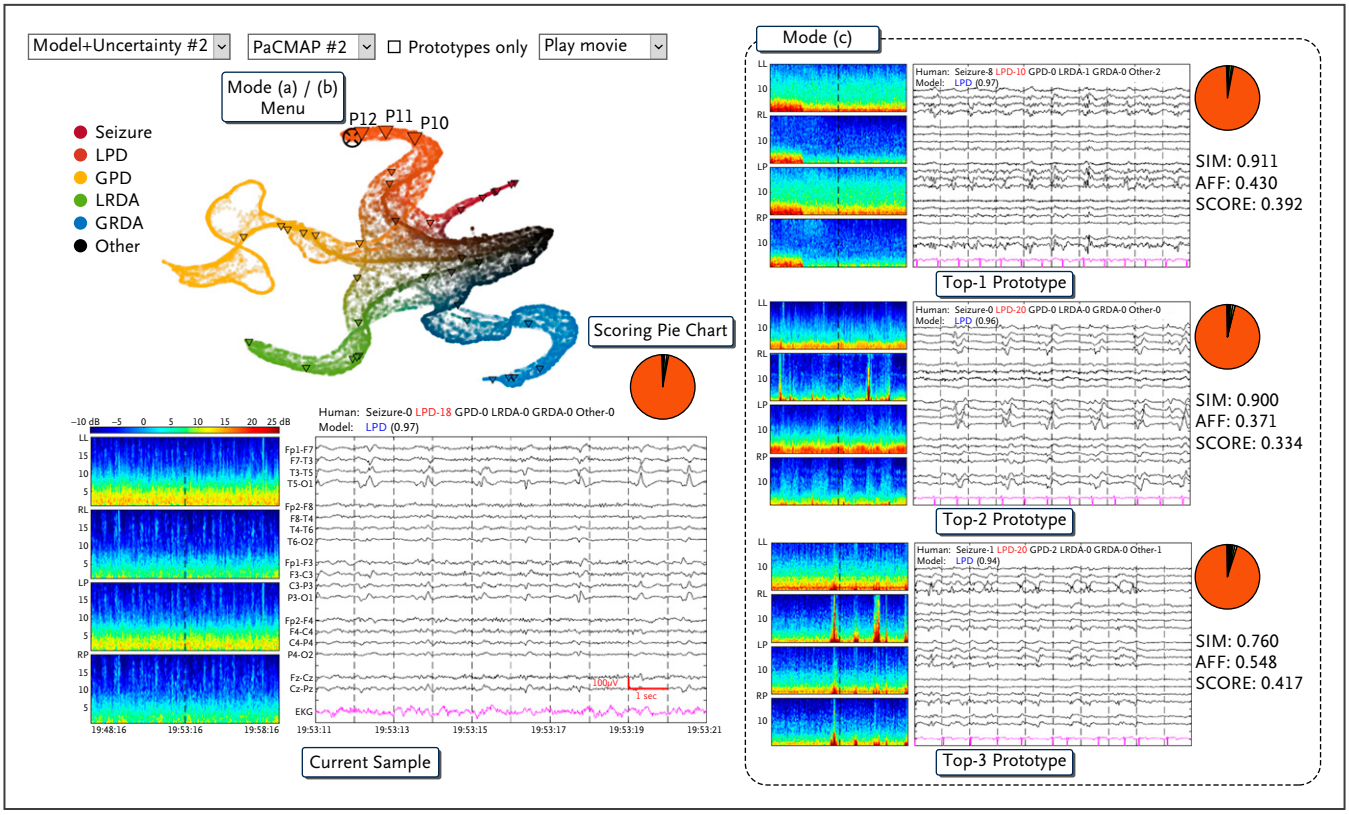


Figure 2. Snapshot of the GUI of the Interpretable System.

The GUI integrates the three explanation modes detailed in Figure 1. This snapshot has minor simplifications for ease of reading. Full details and further information on the GUI layout are provided in the Supplementary Appendix (Section C). AFF denotes affinity; GPD, generalized periodic discharge; GRDA, generalized rhythmic delta activity; GUI, graphical user interface; LL, left lateral; LP, left parasagittal; LPD, lateralized periodic discharge; LRDA, lateralized rhythmic delta activity; RL, right lateral; RP, right parasagittal; and SIM, similarity score.

Each participant was provided with basic training materials for identifying the five IIIC EEG patterns. (A special user interface designed for this user study is shown in Figure 3.) The study comprised two stages set 2 weeks apart; in each, users were asked to classify the same 100 samples as one of five IIIC patterns or as other or no idea. The 100 samples were selected from the test set to ensure that all the classes were equally represented, that the expert annotators had a high level of agreement with each other, and that patients did not appear twice within the same class. Users were randomly split into two groups. One group was given AI assistance in only the first stage, and the other group was given AI assistance in only the second stage. After the users completed both stages, they were asked to complete a survey with questions regarding the study.

### EXTERNAL VALIDATION

To validate model performance across institutions, we collected a new dataset of 1500 events from 327 patients in the intensive care unit (ICU) who underwent continuous EEG as part of clinical care at Brigham and Women's Hospital. Following the same instructions as for labelling the original dataset, 10 EEG experts labeled the middle 10 seconds of 50-second EEG segments, and raters produced one of the six labels.

## Results

### MODEL CLASSIFICATION PERFORMANCE

Model performance was evaluated by using area under the receiver-operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC) scores. Both AUROC and AUPRC were calculated by using the predicted class probability output from the softmax layer of the model. The classification performance of our interpretable model ProtoPMed-EEG significantly exceeds that of the uninterpretable state-of-the-art baseline for this task, SPaRCNet,[14] in distinguishing seizures, LPDs, GPDs, LRDAs, and GRDAs as measured both by AUROC and AUPRC scores ($P<0.001$).

Results for receiver-operating characteristic and precision–recall curve analyses are shown in Figure 4. Seizure versus no-seizure classification performance can be found in Figure 4A under the heading seizure. For comparing AUROC scores, the DeLong test[31] was used for statistical significance. For AUPRC comparisons, we tested for statistical significance using the bootstrapping method with 1000 bootstrap samples. These findings held when bootstrapping

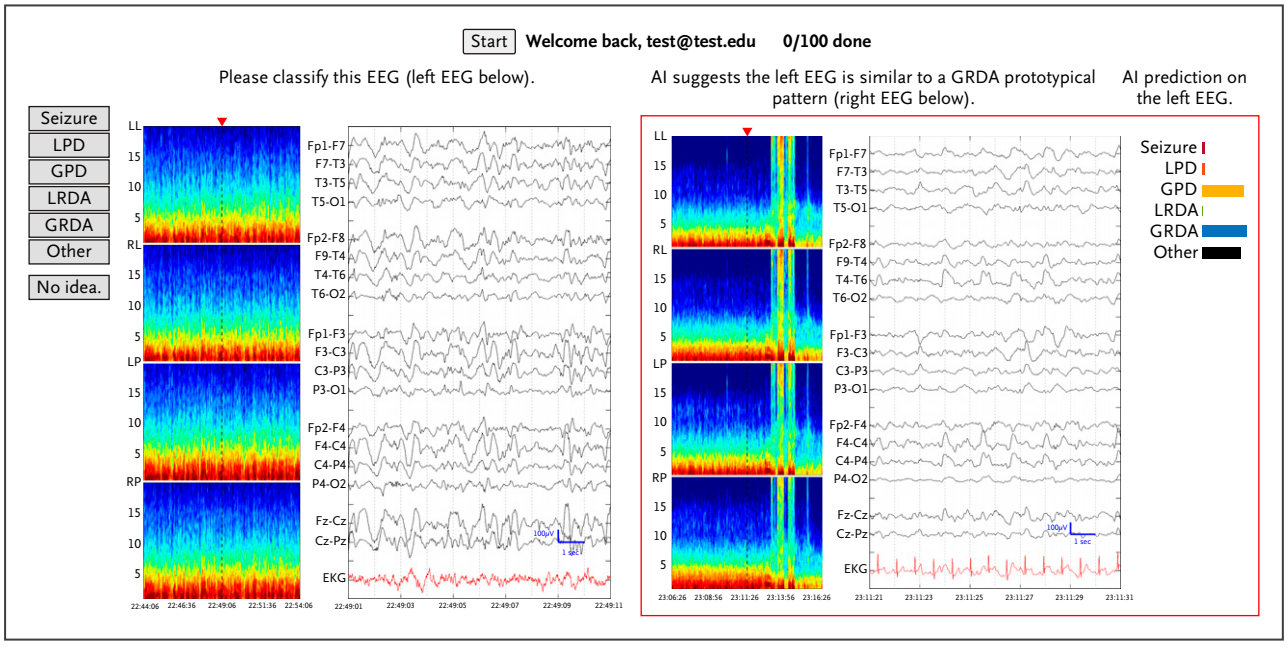


Figure 3. The User Study Interface When AI Assistance Is Provided.

Buttons to select the electroencephalography (EEG) pattern category, the EEG sample to be categorized, a comparable EEG prototype provided by the model, and a bar chart of class predictions from the model are shown from left to right. AI denotes artificial intelligence; EKG, electrocardiogram; GPD, generalized periodic discharge; GRDA, generalized rhythmic delta activity; LL, left lateral; LP, left parasagittal; LPD, lateralized periodic discharge; LRDA, lateralized rhythmic delta activity; RL, right lateral; and RP, right parasagittal.

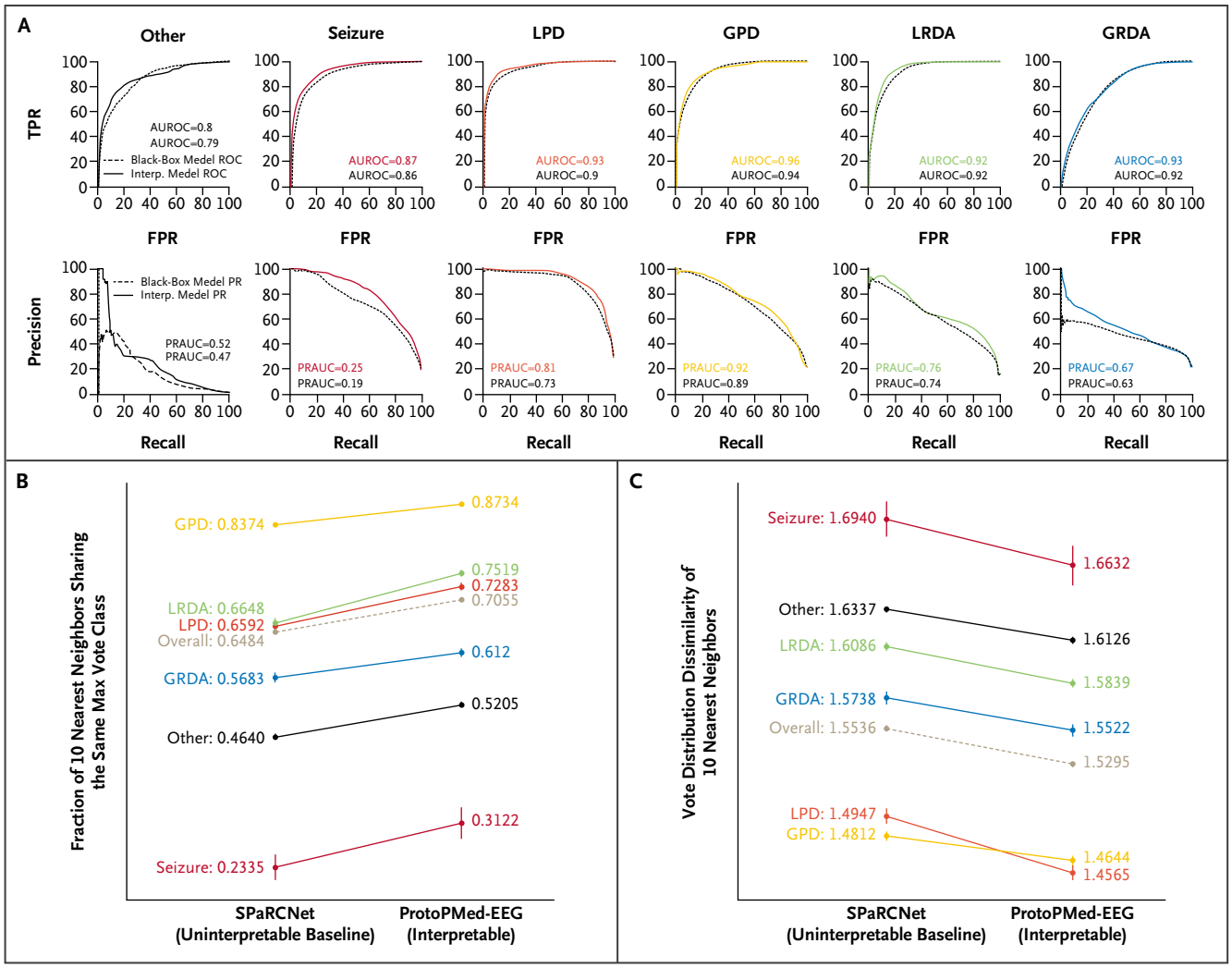


Figure 4. Performance Evaluation.

We compare the area under the receiver-operating characteristic curve (AUROC) scores and area under the precision–recall (PR) curve scores (Panel A), the neighborhood analysis by maximum vote (higher value indicates better clustering by class; uses only majority vote of each sample; Panel B), and the neighborhood analysis by annotator vote distributions (lower value indicates better clustering by class) between the uninterpretable SPaRCNet[14] and our interpretable ProtoPMed-EEG (Panel C). In Panel A, the receiver-operating characteristic (ROC) curves and PR curves for ProtoPMed-EEG (solid lines) are compared with SPaRCNet (dashed lines). ProtoPMed-EEG has statistically significantly higher AUROC and AUPRC. In Panel B, we shown the neighborhood analysis by maximum. In Panel C, we show the neighborhood analysis by annotator vote distribution (lower values mean a more consistent neighborhood). FPR denotes false positive rate; GPD, generalized periodic discharge; GRDA, generalized rhythmic delta activity; Interp., interpretable; LPD, lateralized periodic discharge; LRDA, lateralized rhythmic delta activity; PRAUC, area under the precision–recall curve; and TPR, true positive rate.

according to patient or sample. More details on these significance tests are provided in the Supplementary Appendix (Section E).

## USER STUDY

Of the 13 invited users, 8 completed both stages of the study, 2 completed only the first stage and did not complete the second stage 2 weeks later, and 7 filled out the poststudy survey. Participant dropout rates were similar in both stages.

Mean user accuracy in identifying the correct class was significantly better for all users with AI assistance than without (71 vs. 47%; one-sided Student's t-test, $P<0.05$), as shown in Table S3. Individual results are shown in Figure 5C.

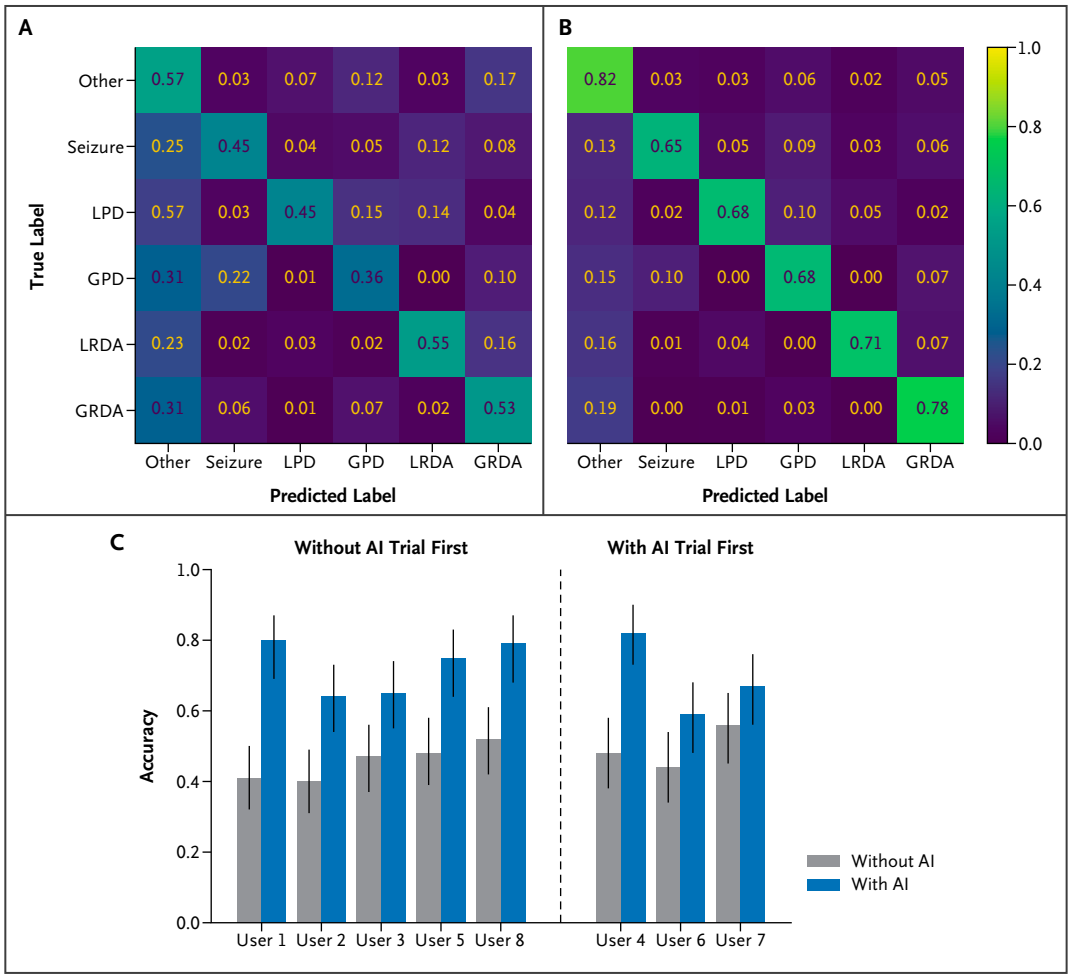


Figure 5. Results of the User Study.

Average error matrix across all users without artificial intelligence (AI; Panel A). Average error matrix across all users with AI (Panel B). Electroencephalography pattern classification performance of the users with and without AI. All users performed significantly better ($P<0.05$) while provided with AI assistance (Panel C). GPD denotes generalized periodic discharge; GRDA, generalized rhythmic delta activity; LPD, lateralized periodic discharge; and LRDA, lateralized rhythmic delta activity.

The average time taken to make a decision was longer with AI assistance than without (32±33 seconds vs. 25±39 seconds). An increase in mean IRR was observed. Rater-to-majority IRR percent agreement was 90% with AI assistance and 86% without; Cohen's kappa scores were 66% with AI assistance and 48% without. Mean rater-to-rater IRR percent agreement was 62% with AI assistance and 47% without; Cohen's kappa scores were 54% with AI assistance and 35% without.

In the poststudy survey, seven of seven users believed their ability to identify EEG patterns improved after completing the stage with AI, while three users felt they improved after the stage without AI. All seven users would recommend this system to medical professionals learning to identify these patterns. Further analysis of the user study results is provided in the Supplementary Appendix.

## EXTERNAL VALIDATION

The model–rater agreement was calculated by using AUROC and AUPRC. On the external dataset, we found an average AUROC of 0.85 and an average AUPRC of 0.61. On the internal dataset, we found an average AUROC of 0.91 and an average AUPRC of 0.74. Our model maintained high predictive performance, despite shifts in class distribution and annotator population. A large shift in class distribution existed between the two datasets, with the other class

comprising 60% of the external dataset but only 22% of the internal dataset. A shift in the population of annotators existed between the internal and external datasets, with 10 annotators for the external dataset and 124 for the internal dataset. Further details are provided in the Supplementary Appendix.

### MODEL INTERPRETABILITY PERFORMANCE

A key part of our explanation is to provide prototypical EEG samples that are comparable with the sample being analyzed. The interpretable model judges two samples' similarity from their distance in the latent space based on key features (Fig. 1A). Because the explanation depends on the meaningful placement of samples in the latent space, we expect the feature extractor to place samples with high resemblance close together. In an ideally structured latent space, we would have neighborhoods of consistent samples. In particular, they would share the same class label and other medically relevant characteristics.

To evaluate the explanatory power of the learned prototypes and the meaningfulness of the latent space, we designed two metrics to measure neighborhood consistency. In Figure 4B, for each sample in the test set, we calculated the fraction of the 10 nearest test set neighbors where the class with the most votes was the same as for the sample (by maximum). In Figure 4C, we considered the neighborhood analyses by vote such that for each sample, we calculated the mean cross-entropy of the vote distribution of the sample with the vote distribution of each of the 10 nearest neighbors (Fig. 4C). Here, we considered cross-entropy as a discrete distribution across classes and checked whether the cross-entropy of the test point matched the distribution of classes from the nearest neighbors. The interpretable model performed significantly better than the black-box model across all metrics and classes (with $P<0.05$ for each comparison). Qualitative assessments of the neighborhoods are provided in the Supplementary Appendix (Section G).

### MAPPING THE IIIC

Our model provides evidence for the concept of a continuum between ictal and interictal EEG patterns; a set of well-separated classes would instead be represented by disconnected islands. This morphology is consistent across models initiated with different random seeds. This perspective is supported by the correlation between our model's predictions and the labeling experts' opinions. Our model successfully identified samples that were categorized as between classes, in which the class probabilities assigned by the model closely matched the distribution of expert votes (i.e., split across two or more classes). In Figure 1C, the coloring and distance between samples (points) were based on the model class scores for each sample. This resulted in a structure with outer points (arms) corresponding to single classes and revealed dense, thread-like paths mapping a gradual change between IIIC classes.

We further sampled along paths between each pair of IIIC patterns and produced videos demonstrating the smooth continuum from one pattern to the other. Videos are provided at https://warpwire.duke.edu/w/8zoHAA/.

## Discussion

In this study, we developed the first inherently interpretable deep-learning model to classify IIIC activity. A user study with first-of-its-kind AI assistance using case-based explanations was included. We showed that when users are provided with interpretable AI assistance, their accuracy in predicting IIIC EEG patterns significantly improves, demonstrating the efficacy of this system for human–AI collaboration. We also showed that the model could generalize to a dataset from another hospital. Compared with the current state-of-the-art black-box models for this task, the interpretable model achieved better neighborhood analysis scores, indicating that it learned purer neighborhoods in the latent space; that is, the geometry of our latent space, which groups samples from the same class, forms neighborhoods without many samples from outside the class. This method is useful for providing related EEG samples as part of its explanations. Our work thus yields advances in terms of the model's predictive performance and interpretability.

Machine learning, and specifically deep learning, has been used for EEG classification tasks, including seizure detection, with satisfactory predictive performance. Previous studies have produced fully automated black-box models[14,32,33] and black-box models with post hoc explanations[21-25] to address interpretability challenges. However, at best, post hoc methods only approximate model reasoning, and different methods will generate conflicting explanations. In many cases, it is guaranteed that the explanation will not match true reasoning. In contrast, our explanation follows the exact path within the model as the prediction generation, thus offering perfect explanation faithfulness.

Our interpretable model goes beyond the automated detection in black-box models, providing clinicians with

the means to validate diagnoses. The explanations produced by our model include a graphical representation of the sample's relative position to all learned prototypical samples along the underlying IIIC, visual comparisons with relevant samples, and an easy-to-understand scoring system based on the learned similarity to the prototypical samples. These explanation components help users gauge the appropriate level of trust for a specific prediction based on the model's explicit reasoning.

Our model is not only novel in its applications to neurology, but it also provides substantial improvements to the existing interpretable prototype-based neural network literature. Although interpretable deep-learning models are available for medical applications, most are limited to the computer vision domain. Barnett et al.[34] provide an inherently interpretable system for a breast mass classification task, but it is limited to computer vision applications using prototypes that represent one part of an image. In past work on leveraging prototypes to provide explanations for model predictions,[35-37] each prototype was limited to represent a single class; that setting is insufficient for mapping IIIC EEG signals because some present defining features of two classes.

Before our study, seizure and seizure-like patterns were treated as isolated classes; however, in reality, they form a continuous space as proposed in the IIIC hypothesis. Our introduction of dual-class prototypes enables our model to place prototypes between two classes in the latent space, providing insights into EEG patterns in transitional states. In addition, the inherent interpretability of our model would facilitate adoption of this deep-learning model in real-world practice as it provides humans with adequate visibility into the model's reasoning process to reduce potential misdiagnoses.

The ProtoPMed-EEG model's expert-level predictive performance and its interpretable nature make it a promising candidate for application in clinical ICU settings. The enhancements showcased by participants in the user study, in which a purpose-built GUI incorporates the model's explanations, highlight the model's potential to mitigate human subjectivity and enhance user classification accuracy, particularly for challenging IIIC EEG patterns. Specifically, this technology can be used to enhance a neurologist's consultation, to act as a neurologist consultation when a neurologist is not available, or to help determine whether consultation with a neurologist is needed for a patient. Particularly for hard-to-diagnose seizure-like EEG patterns, the explanations given by the model can provide relevant examples for a neurologist during the diagnostic step. In scenarios in which a neurologist is not readily available (e.g., settings that lack 24/7 neurologist coverage), the system can serve as an interim consultant; nonspecialist clinicians (e.g., ICU nurses, ICU residents, non-neurology ICU physicians) could compare the EEG readings at hand with expert-annotated prototypes provided by the model to reach an informed decision. Such a system could also assist ICU non-neurologists in triaging which episodes need further consultation with a neurologist. In addition, with little to no modifications, this system presents a low-cost and interactive training tool for physicians and clinical practitioners for IIIC EEG pattern recognition and classification.

Our TEEGLLTEEG approach of allowing the best-matching prototype to be in-between classes (dual-class prototype) can inform disease phenotyping for neurologic conditions. This information is valuable because the potential for harm to the brain depends not only on the class of IIIC activity but also on additional characteristics such as the frequency of discharges (e.g., LPDs at 2 Hz are worse than LPDs at 1 Hz). Also, LPD patterns at lower frequencies typically match single-class prototypes, whereas LPDs that match dual-class seizure–LPD prototypes tend to be LPDs at higher frequencies or to have other features that make them more like seizures. Matching to dual-class prototypes as opposed to simply declaring the class label provides an opportunity for clinicians to take account of these additional features. This feature-level interpretability, in turn, may inform clinical decision-making by allowing treating clinicians to better match treatment intensity and risk to the potential for harm posed by a given IIIC pattern: for example, to select oral or nonanesthetic antiseizure medications for less harmful IIIC patterns while reserving more aggressive intravenous anesthetic drips for more harmful IIIC patterns.

There are limitations to the current study. First, we used the majority vote of the 124 raters as ground truth. Even though the rater population was large, it is still possible that a different group of raters might yield different ground truth labels. Second, the interpretability is limited to the final steps of the model, and thus a clinician must still infer how the relevant qualities (i.e., peak-to-peak distance, amplitude, burst suppression ratio) are weighed in the model's assessment of similarity. In cases in which this not clear, qualitative neighborhood analyses are available for examination. Future work could account for known qualities of interest explicitly. Despite this limitation, the

interpretability provided by this model greatly surpasses that of the state of the art for this task.

## Conclusions

We developed an interpretable deep-learning algorithm that accurately classifies six clinically relevant EEG patterns and offers faithful explanations for its classifications by leveraging prototype learning. We show that users have superior classification performance when provided with this AI assistance. The comprehensive explanations and the GUI enable follow-up user studies of clinical applications, including diagnostic assistance and education.

### Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Supported by the National Science Foundation (grant numbers IIS-2147061 [with Amazon], HRD-2222336, IIS-2130250, and 2014431) and the National Institutes of Health (grant numbers R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, and K23NS124656). Dr. Taraschenko received support from the National Institutes of Health (grant number P20GM130447, Cognitive Neuroscience and Development of Aging Award) and the Nebraska Department of Health and Human Services (grant number LB606, Stem Cell Grant).

We acknowledge Drs. Aaron F. Struck, Safoora Fatima, Aline Herlopian, Ioannis Karakis, Jonathan J. Halford, Marcus Ng, Emily L. Johnson, Brian Appavu, Rani A. Sarkis, Gamaleldin Osman, Peter W. Kaplan, Monica B. Dhakar, Lakshman, Arcot Jayagopal, Zubeda Sheikh, Olha Taraschenko, Sarah Schmitt, Hiba A. Haider, Jennifer A. Kim, Christa B. Swisher, Nicolas Gaspard, Mackenzie C. Cervenka, Andres Rodriguez, Jong Woo Lee, Mohammad Tabaeizadeh, Emily J. Gilmore, Kristy Nordstrom, Ji Yeoun Yoo, Manisha Holmes, Susan T. Herman, Jennifer A. Williams, Jay Pathmanathan, Fábio A. Nascimento, Mouhsin M. Shafi, Sydney S. Cash, Daniel B. Hoch, Andrew J. Cole, Eric S. Rosenthal, Sahar F. Zafar, and Jimeng Sun, who played major roles in creating the labeled EEG dataset and SPaRCNet used in the study. We also acknowledge Rachel Choi, Justin Sattin, Kayla N. Haffley, Anthony P. Tran, Ushna Khan, Luana Rodrigues Dos Santos, HyoJin Park, and Emma Locke for participating in our user study. In addition, we thank Piotr Suder for lending his statistical expertise.

### Author Affiliations

[1] Computer Science, Duke University, Durham, NC
[2] Pratt School of Engineering, Duke University, Durham, NC
[3] Beth Israel Deaconess Medical Center, Harvard University, Cambridge, MA
[4] Johns Hopkins Medicine, Baltimore
[5] Emory University, Atlanta
[6] Yale University, New Haven, CT
[7] University of Nebraska Medical Center, Omaha
[8] University of Mississippi Medical Center, Jackson
[9] Mayo Clinic, Rochester, MN
[10] Duke University, Durham, NC

### References

1. Towne AR, Waterhouse EJ, Boggs JG, et al. Prevalence of nonconvulsive status epilepticus in comatose patients. Neurology 2000;54:340-345. DOI: 10.1212/WNL.54.2.340.
2. Jordan KG. Nonconvulsive status epilepticus in acute brain injury. J Clin Neurophysiol 1999;16:332-340. DOI: 10.1097/00004691-199907000-00005.
3. De Marchis GM, Pugin D, Meyers E, et al. Seizure burden in subarachnoid hemorrhage associated with functional and cognitive outcome. Neurology 2016;86:253-260. DOI: 10.1212/WNL.0000000000002281.
4. Payne ET, Zhao XY, Frndova H, et al. Seizure burden is independently associated with short term outcome in critically ill children. Brain 2014;137:1429-1438. DOI: 10.1093/brain/awu042.
5. Lee JW, LaRoche S, Choi H, et al. Development and feasibility testing of a critical care EEG monitoring database for standardized clinical reporting and multicenter collaborative research. J Clin Neurophysiol 2016;33:133-140. DOI: 10.1097/WNP.0000000000000230.
6. Zafar SF, Rosenthal ES, Jing J, et al. Automated annotation of epileptiform burden and its association with outcomes. Ann Neurol 2021;90:300-311. DOI: 10.1002/ana.26161.
7. Parikh H, Hoffman K, Sun H, et al. Effects of epileptiform activity on discharge outcome in critically ill patients in the USA: a retrospective cross-sectional study. Lancet Digit Health 2023;5:e495-e502. DOI: 10.1016/S2589-7500(23)00088-2.
8. Pohlmann-Eden B, Hoch DB, Cochius JI, Chiappa KH. Periodic lateralized epileptiform discharges — a critical review. J Clin Neurophysiol 1996;13:519-530. DOI: 10.1097/00004691-199611000-00007.
9. Rubinos C, Reynolds AS, Claassen J. The ictal-interictal continuum: to treat or not to treat (and how)? Neurocrit Care 2018;29:3-8. DOI: 10.1007/s12028-017-0477-5.
10. Nafea MS, Ismail ZH. Supervised machine learning and deep learning techniques for epileptic seizure recognition using EEG signals — a systematic literature review. Bioengineering (Basel) 2022;9:781. DOI: 10.3390/bioengineering9120781.
11. Gramacki A, Gramacki J. A deep learning framework for epileptic seizure detection based on neonatal EEG signals. Sci Rep 2022;12:13010. DOI: 10.1038/s41598-022-15830-2.
12. Ahmad I, Wang X, Zhu M, et al. EEG-based epileptic seizure detection via machine/deep learning approaches: a systematic review [retracted]. Comput Intell Neurosci 2022;2022:6486570. DOI: 10.1155/2022/6486570.
13. Abdelhameed A, Bayoumi M. A deep learning approach for automatic seizure detection in children with epilepsy. Front Comput Neurosci 2021;15:650050. DOI: 10.3389/fncom.2021.650050.
14. Jing J, Ge W, Hong S, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG

interpretation. Neurology 2023;100:e1750-e1762. DOI: 10.1212/WNL.0000000000207127.

15. Ge W, Jing J, An S, et al. Deep active learning for interictal ictal injury continuum EEG patterns. J Neurosci Methods 2021;351:108966. DOI: 10.1016/j.jneumeth.2020.108966.

16. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: CHI '20: proceedings of the 2020 CHI conference on human factors in computing systems. New York, NY: Association for Computing Machinery, 2020:1-12. DOI: 10.1145/3313831.3376718.

17. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15:e1002683. DOI: 10.1371/journal.pmed.1002683.

18. U.S. Food and Drug Administration. Good machine learning practice for medical device development: guiding principles. U.S. FDA, Health Canada, Medicines and Healthcare Products Regulatory Agency. October 27, 2021 (https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles).

19. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation." AI Mag 2017;38:50-57. DOI: 10.1609/aimag.v38i3.2741.

20. Hamon R, Junklewitz H, Sanchez I, et al. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. IEEE Comput Intell Mag 2022;17:72-85. DOI: 10.1109/MCI.2021.3129960.

21. Zhao X, Yoshida N, Ueda T, Sugano H, Tanaka T. Epileptic seizure detection by using interpretable machine learning models. J Neural Eng 2023;20:015002. DOI: 10.1088/1741-2552/acb089.

22. Uyttenhove T, Maes A, Van Steenkiste T, et al. Interpretable epilepsy detection in routine, interictal EEG data using deep learning. In: Alsentzer E, McDermott MBA, Falck F, Sarkar SK, Roy S, Hyland SL, eds. Proceedings of the machine learning for health NeurIPS workshop, volume 136 of proceedings of machine learning research. 2020:355-366. https://proceedings.mlr.press/v136/uyttenhove20a.html.

23. Jemal I, Mezghani N, Abou-Abbas L, et al. An interpretable deep learning classifier for epileptic seizure prediction using EEG data. IEEE Access 2022;10:60141-60150. DOI: 10.1109/ACCESS.2022.3176367.

24. Gabeff V, Teijeiro T, Zapater M, et al. Interpreting deep learning models for epileptic seizure detection on EEG signals. Artif Intell Med 2021;117:102084. DOI: 10.1016/j.artmed.2021.102084.

25. Lo Giudice M, Varone G, Ieracitano C, et al. Permutation entropy-based interpretability of convolutional neural network models for interictal EEG discrimination of subjects with epileptic seizures vs. psychogenic non-epileptic seizures. Entropy (Basel) 2022;24:102. DOI: 10.3390/e24010102.

26. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206-215. DOI: 10.1038/s42256-019-0048-x.

27. Ng MC, Gaspard N, Cole AJ, et al. The standardization debate: a conflation trap in critical care electroencephalography. Seizure 2015;24:52-58. DOI: 10.1016/j.seizure.2014.09.017.

28. Jing J, Ge W, Struck AF, et al. Interrater reliability of expert electroencephalographers identifying seizures and rhythmic and periodic patterns in EEGs. Neurology 2023;100:e1737-e1749. DOI: 10.1212/WNL.0000000000201670.

29. Chen C, Li O, Tao D, et al. This looks like that: deep learning for interpretable image recognition. Advances in Neural Information Processing Systems 32 (NeurIPS 2019). 2019:8930-8941. DOI: 10.48550/arXiv.1806.10574.

30. Donnelly J, Barnett AJ, Chen C. Deformable protopnet: an interpretable image classifier using deformable prototypes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022:10255-10265. DOI: 10.1109/CVPR52688.2022.01002.

31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837-845. DOI: 10.2307/2531595.

32. Tzallas AT, Tsipouras MG, Tsalikakis DG, et al. Automated epileptic seizure detection methods: a review study. In: Stevanovic D, ed. Epilepsy — histological, electroencephalographic and psychological aspects. London: IntechOpen Limited, 2012:2027-2036.

33. Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 2019;16:031001. DOI: 10.1088/1741-2552/ab0ab5.

34. Barnett AJ, Schwartz FR, Tao C, et al. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nat Mach Intell 2021;3:1061-1070. DOI: 10.1038/s42256-021-00423-x.

35. Zhang X, Gao Y, Lin J, et al. Tapnet: multivariate time series classification with attentional prototypical network. Proc Conf AAAI Artif Intell 2024;34:6845-6852. DOI: 10.1069/aaai.v34i04.6165.

36. Huang C, Wu X, Zhang X, et al. Deep prototypical networks for imbalanced time series classification under data scarcity. In: Proceedings of the 28th ACM international conference on information and knowledge management. 2019:2141-2144. DOI: 10.1145/3357384.3358162.

37. Gee AH, Garcia-Olano D, Ghosh J, Paydarfar D. Explaining deep classification of time-series data with learned prototypes. CEUR Workshop Proc 2019;2429:15-22. DOI: 10.48550/arXiv.1904.08935.