Published in final edited form as:

NEJM AI. 2025 July ; 2(7): . doi:10.1056/aioa2401221.

Expert-Level Detection of Epilepsy Markers in EEG on Short and Long Timescales

J. Li[1,2,3], **D. M. Goldenholz**[2,3], **M. Alkofer**[2,3,4], **C. Sun**[2,3], **F. A. Nascimento**[5], **J. J. Halford**[6,7], **B. C. Dean**[8], **M. Galanti**[8,9], **A. F. Struck**[5,10], **A. S. Greenblatt**[5], **A. D. Lam**[2,11], **A. Herlopian**[12], **C. Nwankwo**[13], **D. Weber**[14], **D. Maus**[2,11], **H. A. Haider**[15,16], **I. Karakis**[17,18], **J. Y. Yoo**[19], **M. C. Ng**[20], **O. Selioutski**[21,22], **O. Taraschenko**[23], **G. Osman**[24], **R. Katyal**[25], **S. E. Schmitt**[6,26], **S. Benbadis**[27,28], **S. S. Cash**[2,11], **W. O. Tatum**[24], **Z. Sheikh**[29], **W. Y. Kong**[2,3], **G. Bayas**[2,3], **N. Turley**[2,3], **S. Hong**[1], **M. B. Westover**[2,3], **J. Jing**[2,3]

[1]National Institute of Health Data Science, Peking University, Beijing

[2]Harvard Medical School, Boston

[3]Neurology Department, Beth Israel Deaconess Medical Center, Boston

[4]Institute for Theoretical Physics, Technical University Berlin, Berlin

[5]Neurology Department, Washington University in St. Louis, St Louis, MO

[6]Ralph H. Johnson VA Medical Center, Charleston, SC

[7]Electrical and Computer Engineering Department, Clemson University, Clemson, SC

[8]Clemson University School of Computing, Clemson, SC

[9]Public Health Sciences Department, Medical University of South Carolina, Charleston

[10]University of Wisconsin–Madison, Madison

[11]Neurology Department, Massachusetts General Hospital, Boston

[12]Yale University School of Medicine, New Haven, CT

[13]Akron Children's Hospital, Akron, OH

[14]St. Louis University School of Medicine, St Louis, MO

[15]Neurology Department, University of Chicago, Chicago

[16]University of Chicago Medical Center, Chicago

Dr. Jing can be contacted at jjing@bidmc.harvard.edu or at Beth Israel Deaconess Medical Center, Neurology Department, 330 Brookline Ave, Boston, MA 02215.

The authors' full names and academic degrees are as follows: Jun Li, B.Eng., Daniel M. Goldenholz, M.D., Ph.D., Moritz Alkofer, M. Sc., Chenxi Sun, Ph.D., Fabio A. Nascimento, M.D., Jonathan J. Halford, M.D., Brian C. Dean, Ph.D., Mattia Galanti, M.Sc., Aaron F. Struck, M.D., Adam S. Greenblatt, M.D., Alice D. Lam, M.D., Ph.D., Aline Herlopian, M.D., Chinasa Nwankwo, M.D., Dan Weber, D.O., Douglas Maus, M.D., Ph.D., Hiba A. Haider, M.D., Ioannis Karakis, M.D., Ph.D., Ji Yeoun Yoo, M.D., Marcus C. Ng, M.D., Olga Selioutski, D.O., Olga Taraschenko, M.D., Ph.D., Gamaleldin Osman, M.D., Roohi Katyal, M.B.B.S., Sarah E. Schmitt, M.D., Selim Benbadis, M.D., Sydney S. Cash, M.D., Ph.D., William O. Tatum, D.O., Zubeda Sheikh, M.D., Wan Yee Kong, M.D., M.Sc., Grace Bayas, B.Sc., Niels Turley, B.Sc., Shenda Hong, Ph.D., M. Brandon Westover, M.D., Ph.D., and Jin Jing, Ph.D.

Mr. Li, Dr. Goldenholz, and Mr. Alkofer and Drs. Hong, Westover, and Jing contributed equally to this article.

The authors' full names and academic degrees are listed in the Appendix.

[17]Emory University School of Medicine, Atlanta

[18]University of Crete School of Medicine, Heraklion, Greece

[19]Icahn School of Medicine at Mount Sinai, New York, NY

[20]University of Manitoba, Winnipeg, MB, Canada

[21]Stony Brook University, Stony Brook, NY

[22]University of Rochester, Rochester, NY

[23]University of Nebraska Medical Center, Omaha

[24]Mayo Clinic, Jacksonville, FL

[25]Louisiana State University Health Shreveport, Shreveport

[26]Neurology Department, Medical University of South Carolina, Charleston

[27]University of South Florida, Tampa

[28]Tampa General Hospital, Tampa, FL

[29]Neurology Department, Virginia Commonwealth University, Richmond

Abstract

BACKGROUND: Epileptiform discharges, or spikes, within electroencephalogram (EEG) recordings are essential for diagnosing epilepsy and localizing seizure origins. Artificial intelligence (AI) offers a promising approach to automating detection, but current models are often hindered by artifact-related false positives and often target either event- or EEG-level classification, thus limiting clinical utility.

METHODS: We developed SpikeNet2, a deep-learning model based on a residual network architecture, and enhanced it with hard-negative mining to reduce false positives. Our study analyzed 17,812 EEG recordings from 13,523 patients across multiple institutions, including Massachusetts General Brigham (MGB) hospitals. Data from the Human Epilepsy Project (HEP) and SCORE-AI (SAI) were also included. A total of 32,433 event-level samples, labeled by experts, were used for training and evaluation. Performance was assessed using the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), calibration error, and a modified area under the curve (mAUC) metric. The model's generalizability was evaluated using external datasets.

RESULTS: SpikeNet2 demonstrated strong performance in event-level spike detection, achieving an AUROC of 0.973 and an AUPRC of 0.995, with 44% of experts surpassing the model on the MGB dataset. In external validation, the model achieved an AUROC of 0.942 and an AUPRC of 0.948 on the HEP dataset. For EEG-level classification, SpikeNet2 recorded an AUROC of 0.958 and an AUPRC of 0.959 on the MGB dataset, an AUROC of 0.888 and an AUPRC of 0.823 on the HEP dataset, and an AUROC of 0.995 and an AUPRC of 0.991 on the SAI dataset, with 32% of experts outperforming the model. The false-positive rate was reduced to an average of nine spikes per hour.

CONCLUSIONS: SpikeNet2 offers expert-level accuracy in both event-level spike detection and EEG-level classification, while significantly reducing false positives. Its dual functionality and robust performance across diverse datasets make it a promising tool for clinical and telemedicine applications, particularly in resource-limited settings. (Funded by the National Institutes of Health and others.)

Introduction

Electroencephalography (EEG) is the primary diagnostic tool for evaluating patients with suspected epilepsy.[1–3] Epileptiform discharges, characterized by distinctive spiky or sharp wave morphologies,[4] serve as fundamental biomarkers of epilepsy and cortical hyperexcitability.[5] These discharges, commonly referred to as spikes, can manifest sporadically (interictal epileptiform discharges), in periodic patterns, or as components of more complex EEG patterns, such as seizures. In clinical practice, the detection of spikes relies on visual interpretation of EEGs by physicians with specialized training in clinical neurophysiology. These interpretations are crucial for making therapeutic decisions and distinguishing epilepsy from other conditions that mimic seizures.[6] Despite the fact that epilepsy affects over 70 million people globally,[7] in many parts of the world, EEGs are interpreted by physicians without specialty training[8,9] — even in regions with adequate general medical care — leading to frequent misinterpretations and misdiagnoses.[10,11] Recent studies have highlighted the importance of spikes beyond epilepsy diagnosis. They serve as indicators of impending delayed cerebral ischemia after subarachnoid hemorrhage,[12] markers for posttraumatic epilepsy risk,[13] and potential contributors to cognitive deterioration in Alzheimer's disease.[14] However, the clinical utility of these findings is constrained by the need for expert manual interpretation of EEGs. Given the increasing value of spike detection and the growing demand for EEG analysis, there is an urgent need for accurate automated methods to detect spikes.[15]

Artificial intelligence (AI) has shown promise in improving epilepsy diagnosis accuracy, providing EEG interpretations where specialized expertise is unavailable, and extending the reach of existing neurologists.[16–19] Two notable large-scale studies on spike detection have resulted in the models SpikeNet1 and SCORE-AI (SAI).[20,21] However, these models address limited aspects of spike detection in isolation.

EEG interpretation for spikes involves two distinct but related tasks: event-level detection (identifying individual spike events) and EEG-level classification (determining whether or not an EEG contains spikes). SpikeNet1, developed using 9571 EEGs, achieved expert-level performance in event-level spike detection on its internal test set. However, it was not evaluated on continuous EEG recordings, which increases the likelihood of false positives, and its EEG-level classification capabilities were limited.[20] SCORE-AI, trained on 30,493 EEGs labeled at the EEG-level, demonstrated expert-level performance for EEG-level classification but did not address event-level spike detection.[21] To date, to our knowledge, no AI model has rigorously demonstrated expert-level spike detection at both event and EEG levels. Moreover, existing methods often lack multicenter training and validation, limiting their demonstrated generalizability.[22] For instance, SpikeNet1 lacked external validation

across multiple centers, while SCORE-AI's training and evaluation were limited to routine EEGs, excluding recordings from critically ill patients.

To address these limitations, we developed SpikeNet2, a residual network (ResNet)–based deep-learning spike-detection model using both event- and EEG-level classification. We implemented a hard-negative mining approach to extract challenging negative training examples from spike-free EEG recordings, significantly reducing false positives.[23] For EEG-level classification, we developed a linear model combining spike-detection probabilities with EEG features, providing a comprehensive epileptiform activity analysis.

We validated SpikeNet2 using a large, multicenter dataset of EEG recordings from patients with and without spikes. Events were independently labeled by 24 expert neurologists, creating, to our knowledge, the largest set of spike samples labeled by multiple experts to date. We also conducted external validation using datasets from the Human Epilepsy Project (HEP) and SAI to assess the model's generalizability across different clinical environments. This rigorous clinical validation of SpikeNet2 used independent test datasets for both event-level and EEG-level classification, and compared SpikeNet2's performance with that of human experts, benchmarked against previously published AI models. Our goal was to demonstrate a model with expert-level performance across levels and datasets.

Methods

An overview of the study is provided in Figure S1 in the Supplementary Appendix.

DESIGN, SETTING, AND PARTICIPANTS

We gathered 17,812 electroencephalograms (EEGs) from 13,523 patients. Model development and validation data characteristics are summarized in Tables 1 and 2 and Table S1. Data sources were:

- Massachusetts General Brigham (MGB) hospitals (17,524 EEGs from 13,235 patients, collected from Massachusetts General Hospital and Brigham and Women's Hospital)
- HEP (188 EEGs, each from a unique patient)
- SAI (100 EEGs, each from a unique patient)

MAIN OUTCOMES AND METRICS

Statistical performance was assessed via the area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), calibration error (Brier score), and the area under a modified receiver operating characteristic curve (mAUC) metric designed to measure sensitivity versus the number of false positives per hour. In addition, two external validation datasets, comprising EEGs from multiple hospitals, including those from HEP and SAI, were used to assess the model's generalizability.

DEVELOPMENT OF THE AI MODEL

Expert Labels: MGB Dataset Event-level labeling of spikes and nonspike events selected from the 17,524 MGB EEGs was performed by 24 experts from 18 different institutions. We selected 32,433 events from 2601 patients' EEGs for labeling, using methods described previously.[17] Eight of the experts were involved in labeling data for SpikeNet1,[20] and 16 were newly recruited for the current project. Each sample was shown to experts within a 15-second window of the surrounding EEG to provide context, mirroring how EEGs are reviewed in clinical practice. The number of votes received per event ranged from 1 to 23, with a median of 4 votes.

TRAINING DATA

Event-Level Classification: We divided the 32,433 multiexpert-labeled MGB events into development data and holdout test sets in a ratio of 9:1. This yielded 29,333 labeled events from 2499 patients for model development. We created two additional subsets from the MGB EEGs to use in model development.

- Weakly labeled samples, comprising 708,877 events from 13,184 EEGs and 9482 patients, were used as additional training samples. These included:
- Samples not labeled by any expert but morphologically similar to labeled spikes based on clustering;[17] and
- Epileptiform discharges extracted from samples labeled in prior work,[24] as occurring within periodic discharges (lateralized and generalized).
- Negative EEGs and hard-negative–mined samples. We identified EEGs from 4456 unique patients whose EEG reports indicated that no spikes were present. We extracted 52,836 nonspike events from 3988 patients through hard-negative mining, a process of identifying challenging negative training samples. The remaining 468 EEGs were used as a holdout test set in calculating the false-positive rate for mAUCs.

EEG-Level Classification: For training the EEG-level classifier, we prepared a set of 4087 patients' EEGs from MGB, with 708 positive EEGs (containing spikes) and 3379 negative EEGs (no spikes) based on experts' consensus and EEG reports.

A detailed Consolidated Standards of Reporting Trials (CONSORT)–like flow diagram illustrating the data splits is shown in Figure 1. In the development and training of the model, the same EEG may be used across different phases. For example, a single negative EEG can contribute both to hard mining events and to the EEG-level control cases. As a result, the total number of EEGs or patients is smaller than the sum of those from each individual phase. In other words, patients and EEGs may overlap across different training phases during model development. Nevertheless, there is strictly no overlap between the training and test sets. For a more detailed description of the process for allocating the data for model development and testing at the patient level, event level, and EEG level, see Supplementary Appendix, Section 1.0.

MODEL TRAINING FOR EVENT-LEVEL CLASSIFICATION

We preprocessed 19-channel referential EEG signals by resampling to 128 Hz and converting them into both 18-channel bipolar and 19-channel common average references, resulting in a 37-channel input. This allowed the model to access both local and global spatial information. EEGs were segmented into 1-second windows (37×128 matrices), suitable for temporal and spatial feature extraction. To improve model generalization and robustness, we applied data-augmentation techniques during training: channel flipping (mirroring hemispheres), cutting (removing segments), and jittering (adding noise). These augmentations were applied only during training. EEG amplitudes were normalized based on the 95th percentile to reduce the influence of outliers and stabilize training.[25] We then trained a convolutional neural network based on a ResNet architecture, incorporating bottleneck residual blocks and channel-wise attention mechanisms (Fig. S2).[26,27] These design choices allowed the model to focus on salient features while controlling complexity. Shortcut connections were included to preserve gradient flow. Given the severe class imbalance between spike and nonspike samples, we used the focal loss function to focus the model on complex examples during training.[28]

Hard-Negative Mining: High false-positive rates are a key limitation for clinical adoption of spike detectors. Our previous model, SpikeNet1, trained on negative EEG segments, did not explicitly identify hard negatives — nonspike patterns resembling spikes. To address this, we implemented hard-negative mining as shown in Figure 2. We first trained a model (SpikeNet2a) on the labeled dataset. We then applied this model to a large set of control EEGs that contained no spike events (as confirmed by clinical reports). We segmented these EEGs using a 1-second window and 0.25-second stride. Any segment for which the model output exceeded a threshold was considered a false positive and labeled as a negative sample.[29] In the first round, we used a threshold of 0.8 to capture high-confidence false positives. These were added to the training set, and the model was retrained. In subsequent rounds, we progressively lowered the threshold, increasing the diversity of hard negatives. After 11 rounds of this process, we collected 52,836 hard-negative samples. The resulting model, trained with these hard negatives, is referred to as SpikeNet2b, or simply SpikeNet2. This iterative process improved the model's ability to suppress artifact-induced false positives while preserving sensitivity to true spikes. Performance improvements are shown in Figure S3.

MODEL TRAINING FOR EEG-LEVEL CLASSIFICATION

While detecting individual spikes is important, clinical interpretation requires determining whether or not the EEG as a whole contains any spikes. To perform EEG-level classification, we applied both SpikeNet1 and SpikeNet2 to generate output probabilities across each EEG recording. Using both models helped capture complementary patterns observed during development. From each time series, we extracted 79 statistical features, including summary statistics (mean, maximum, standard deviation, median, interquartile range), percentiles (from 2.5th to 97.5th), area under the prediction curve, and spike rates at threshold levels from 0.5 to 1 in 0.01 steps. We used these as input to a logistic-regression model with L1 regularization (also known as least absolute shrinkage and selection operator [LASSO]).

This simple but effective model allowed for interpretable integration of temporal-spike evidence.

VALIDATION OF THE AI MODEL

Test Datasets: We validated SpikeNet2 using three independent datasets that were not involved in model development. These datasets included both event-level and EEG-level labels provided by clinical experts. All testing was conducted using a frozen model. Based on sample-size calculations for ROC curve analysis,[30] a minimum of 60 EEGs was needed to ensure high-confidence AUC estimates; all datasets exceeded this.

External Validation Dataset 1: SCORE-AI. SAI[21] includes 100 routine EEGs (excluding intensive care unit EEGs) from 11 institutions (Table 2). Each EEG was labeled independently by up to 14 blinded experts. EEGs were defined as: positive, greater than or equal to 75% "yes" votes; negative, 0 "yes" votes; and ambiguous (in-between), excluded. This yielded 65 EEGs (25 positive, 40 negative) for EEG-level validation. Event-level labels were not available.

External Validation Dataset 2: Human Epilepsy Project. HEP includes 143 EEGs from adults with newly diagnosed focal epilepsy. Each EEG was reviewed by three experts. Events were labeled positive if two or more experts voted yes, and ambiguous events were excluded. Negative segments were randomly selected from regions without any expert labels. In total, 3296 labeled events (1650 positive, 1646 negative) were included. EEG-level consensus labels were also used.

Large Single-Center Test Dataset. This internal dataset includes 24,585 events, with the test split containing 1671 events from 216 patients. Events were labeled positive with six or more expert votes out of eight. For EEG-level testing, we used 1000 EEGs (500 positive, 500 negative) based on clinical EEG reports, which were reviewed by a fellow and attending epileptologist. We also analyzed performance across subgroups including age, ethnicity, spike localization (focal vs. generalized), and neurodevelopmental state (Supplementary Appendix, Section 1.5).

Outcome Measures. Discrimination metrics: We evaluated model performance using AUROC and AUPRC, and for event-level performance, we also computed mAUC, which plots sensitivity against false positives per hour (FP/hour). This metric better reflects real-world performance, especially on continuous EEG recordings in which nonspike activity dominates. FP/hour was calculated using 468 spike-free EEGs.

Calibration metrics: Calibration assesses how well the predicted probabilities match actual outcomes. We binned samples into five ranges (0 to 20%, 20 to 40%, 40 to 60%, 60 to 80%, and 80 to 100%) based on expert consensus and compared model predictions using calibration curves and Brier scores. Lower Brier scores indicate better probabilistic accuracy.

Comparison with experts: SpikeNet2 was compared with human experts in terms of discrimination and calibration. We used ROC and PR curves, comparing model performance

to expert labels across multiple datasets. Specifically, we calculated the metric EOSN2 (the percentage of experts who outperform SpikeNet2's ROC, PR, or calibration).

BENCHMARKING AGAINST PREVIOUSLY PUBLISHED AI MODELS

We compared SpikeNet2 with previous models, including SpikeNet1, the SpikeNet2 without hard-negative mining (SpikeNet2a), and SCORE-AI, to assess its performance relative to other state-of-the-art systems. Performance metrics were reported with 95% confidence intervals, derived through 10,000 rounds of bootstrapping, and are reported using the format: X [Y to Z], with X being the real value and Y and Z being the lower and upper bounds of the confidence interval, respectively.

Results

DATASET CHARACTERISTICS

We analyzed 17,812 EEGs from 13,523 patients across three datasets. The primary dataset from MGB hospitals included 17,524 EEGs from 13,235 patients, with ages ranging from 0 to 100 years (median, 53 years; interquartile range, 43 years); was 47% female; and was recorded in outpatient (58%), epilepsy monitoring unit (6%), and long-term monitoring (34%) settings. External validation was performed using the HEP dataset (188 outpatient EEGs) and the SCORE-AI dataset (100 EEGs from the epilepsy monitoring unit and outpatient settings).

MODEL EVALUATION: EVENT-LEVEL CLASSIFICATION

SpikeNet2 was first evaluated for event-level spike classification on the MGB holdout test set, as shown in Figure 3A to Figure 3C. SpikeNet2 achieved an AUROC of 0.973 (95% confidence interval [CI], 0.961 to 0.982) and an AUPRC of 0.995 (95% CI, 0.993 to 0.997). By contrast, SpikeNet2 without hard-negative mining (SpikeNet2a) had an AUROC of 0.888 (95% CI, 0.856 to 0.917) and an AUPRC of 0.976 (95% CI, 0.966 to 0.984). The previous best model, SpikeNet1, tested on the same data, achieved an AUROC of 0.760 (95% CI, 0.710 to 0.806) and an AUPRC of 0.929 (95% CI, 0.907 to 0.950). SpikeNet2 surpassed the baseline models in discrimination, with DeLong's test P values of 4.9E-7 for SpikeNet1 versus SpikeNet2a, 4.1E-14 for SpikeNet1 versus SpikeNet2b, and 7.0E-7 for SpikeNet2a versus SpikeNet2b. The calibration error for SpikeNet2 was 0.02 (95% CI, 0.01 to 0.03), outperforming SpikeNet1 and SpikeNet2a. SpikeNet2 achieved a mAUC of 0.997 (95% CI, 0.994 to 0.998), as shown in Figure 3G. The mAUC initial SpikeNet2a was 0.967 (95% CI, 0.961 to 0.973). SpikeNet1's mAUC was 0.990 (95% CI, 0.987 to 0.992). SpikeNet2 outperformed both baseline models in calibration and false-positive measures. Additional analyses for performance differences with age, ethnicity, spike localization, and neurodevelopmental state show that, with the exception of a few subgroups with insufficient samples to draw statistically significant conclusions, the model demonstrates comparable performance across subgroups, with overlapping confidence intervals (see Supplementary Appendix, Section S1.5, and Figs. S4A to 4H).

SpikeNet2 also outperformed the majority of experts on all metrics, with 44% (95% CI, 25 to 75) of experts' operating points located above SpikeNet2's ROC curve, 6% (95% CI, 6 to 38) above SpikeNet2's PR curve, and no experts (95% CI, 0 to 6) showed better calibration.

On the HEP external holdout dataset, SpikeNet2 achieved an AUROC of 0.942 (95% CI, 0.933 to 0.950) and an AUPRC of 0.948 (95% CI, 0.939 to 0.956), comparable to but slightly underperforming SpikeNet1's AUROC of 0.979 (95% CI, 0.975 to 0.983) and AUPRC of 0.979 (95% CI, 0.974 to 0.983), as shown in Figure 3D and Figure 3E, with DeLong's test P values of 1.0E-98 for SpikeNet1 versus SpikeNet2a, 1.6E-23 for SpikeNet1 versus SpikeNet2b, and 1.4E-56 for SpikeNet2a versus SpikeNet2b. The calibration error for SpikeNet2 was 0.065 (95% CI, 0.053 to 0.078), outperforming SpikeNet1 and SpikeNet2a, as shown in Figure 3F. Note that this result suggests that, consistent with our experience, SpikeNet1 in some cases provides information complementary to SpikeNet2 for event-level spike classification. This observation motivated the use of both models as input to SpikeNet2's EEG-level classification approach (Fig. S5). In addition, we report in Table S2 the false-positive rate and precision (positive predictive value) for each model, across each dataset, at sensitivity (true-positive rate) values of 0.85, 0.90, and 0.95, respectively.

MODEL EVALUATION: EEG-LEVEL CLASSIFICATION

On EEG-level classification, on the internal holdout test set (MGB test dataset), SpikeNet2 achieved an AUROC of 0.958 (95% CI, 0.946 to 0.968) and an AUPRC of 0.959 (95% CI, 0.947 to 0.970), as shown in Figure 4A and Figure 4B. On the HEP external holdout dataset, SpikeNet2 achieved an AUROC of 0.888 (95% CI, 0.829 to 0.941) and an AUPRC of 0.823 (95% CI, 0.712 to 0.911), as shown in Figure 4C and Figure 4D. On the SAI external holdout dataset,[21] SpikeNet2 achieved an AUROC of 0.995 (95% CI, 0.967 to 1) and an AUPRC of 0.991 (95% CI, 0.931 to 1), indistinguishable from the SCORE-AI model, with an AUROC of 0.997 (95% CI, 0.984 to 1) and an AUPRC of 0.995 (95% CI, 0.968 to 1). SpikeNet2 outperformed most experts across all metrics on the SAI dataset, with 32% (95% CI, 0 to 78.6) of expert operating points exceeding SpikeNet2's ROC curve, and 32.4% (95% CI, 0 to 78.6) surpassing its PR curve, as shown in Figure 4F and Figure 4E. Additional investigations into EEG-level classification and alternative choices for the positive and negative cut-off thresholds are presented in Supplementary Appendix, Section S1.6, and Figure S6 to Figure S9.

EXAMPLES OF SPIKE DETECTION

Examples of SpikeNet2's capability to accurately recognize and localize spikes with a variety of morphologies across different brain regions and in various clinical contexts are shown in Figure S10 to Figure S20. Each example is from a unique patient. These figures highlight SpikeNet2's ability to precisely detect single events, not just classify whether a whole EEG or a long period of EEG contains spikes. This temporal precision is important for clinical applications that require identification of the onset and frequency of epileptiform discharges and for tracking trends. The figures also illustrate how SpikeNet2 has learned to ignore sharp transients that are normal EEG features, such as vertex waves (Figs. S19 and S20). This discrimination between epileptiform and nonepileptiform sharp transients, as a

result of the hard-negative mining process during training, helps account for the very low rate of false positives (i.e., an excellent mAUC value) and enhances clinical utility.

Discussion

In this diagnostic accuracy study, we developed and validated a new AI-based spike-detection model, SpikeNet2, which surpasses state-of-the-art systems in both event-level and EEG-level classification of spikes. Our model dramatically reduces false positives for event-level spike detection relative to comparators, while demonstrating expert-level performance on both event- and EEG-level classification, offering a promising solution for clinical EEG interpretation, especially in remote and underserved areas.

SpikeNet2 was trained and validated on a large-scale multi-institutional dataset representing diverse demographic and geographic backgrounds, including patients from Europe, the United States, and Africa. To reduce false positives, we implemented hard-negative mining on long-term EEG data — a novel approach that distinguishes complex EEG patterns (e.g., wicket waves, vertex waves) and other benign variants from epileptiform discharges.[10] The model achieved an average false-positive rate of nine spikes per hour of EEG at 90% sensitivity, a large improvement over the previous state-of-the-art SpikeNet1.[20] This modest false-positive rate enhances the model's precision and minimizes the need for expert double-checking, improving clinical utility in resource-limited settings.[31]

Unlike some proprietary AI-based EEG tools, which do not publicly share training or benchmarking data, both the code (GitHub, https://github.com/bdsp-core/SpikeNet2) and the data (bdsp.io project, https://bdsp.io/content/spikenet/2.0) for SpikeNet2 are freely available, facilitating reproducibility and benchmarking of future models. By fostering an open ecosystem, we aim to promote scientific rigor, transparency, and accelerated development of more accurate and robust AI tools for EEG analysis.

Model development was rigorously designed to address common EEG interpretation errors. Twenty-four experts from 18 institutions annotated the EEG recordings, providing robust labels for model training. Validation included a "fixed and frozen" approach — no adjustments were made during external testing. SpikeNet2 was tested on diverse datasets, including various EEG systems, and blinded analysis was performed to ensure objective performance evaluation. External validation using multiple datasets (i.e., SAI, HEP) confirmed generalizability.

One of SpikeNet2's key innovations is its dual capability — event-level spike detection and EEG-level classification — which provides a more comprehensive analysis than models focused solely on one timescale. Event-level detection enables precise time localization of epileptiform discharges, which is crucial for downstream analyses such as spike-morphology studies,[32–34] localization of seizure foci,[35] and delayed cerebral ischemia risk tracking.[35] At the EEG level, SpikeNet2 provides an overall classification of the presence or absence of epileptiform activity, a vital component of epilepsy diagnosis.

Our model demonstrated accuracy comparable to and exceeding most human experts. In event-level spike detection, SpikeNet2 outperformed 56% of experts in terms of ROC

curve positioning, and 6% of experts exceeded its PR curve performance. In EEG-level classification, SpikeNet2 outperformed 68% of experts across multiple datasets. These results demonstrate SpikeNet2's potential to provide highly reliable EEG readings and enhance diagnostic consistency in clinical settings, overcoming the problem of interrater variability, a known challenge in EEG interpretation performed by human experts.[6,36]

SpikeNet2 also addresses a critical challenge in spike detection in continuous EEG monitoring: false positives. With a mAUC of 0.997 (95% CI, 0.994 to 0.998) for event-level detection, it significantly outperforms SpikeNet1 (mAUC of 0.990 [95% CI, 0.987 to 0.992]). The mAUC metric specifically measures sensitivity versus false positives per hour of EEG, providing a more clinically relevant measure of event-level detection performance for continuous monitoring. SpikeNet2 achieved an average false-positive rate of just 9 spikes per hour of EEG at 90% sensitivity, a substantial improvement on SpikeNet1, which had a false-positive rate of 52 per hour at 90% sensitivity. This dramatic reduction in false positives while maintaining high sensitivity enhances SpikeNet2's clinical utility by minimizing the need for expert review of detected events.

The model has some limitations. Although SpikeNet2 can detect EEG spikes, spikes alone do not confirm epilepsy, nor does their absence rule it out — clinical context and expert interpretation remain essential. SpikeNet2 does not currently address other important aspects of EEG interpretation, such as seizure detection or the analysis of background rhythms like alpha activity. Finally, while SpikeNet2 did not exhibit signs of differences in performance across age, race, ethnicity, or developmental state, our datasets lacked substantial representation of some key demographics, including Asian populations. Further investigation is needed to assess generalizability.

Conclusion

SpikeNet2 is a comprehensive model capable of detecting epileptiform discharges at both the event level and EEG level. Moreover, it was trained and validated on large-scale, demographically diverse datasets. Its superior performance positions it as a valuable tool in clinical practice, particularly in resource-limited settings. As AI advances, we anticipate SpikeNet2's application in telemedicine, especially where experienced neurologists are scarce. The model holds promise for real-time EEG monitoring and seizure diagnosis, paving the way for its integration into clinical workflows. Future work will focus on expanding its capabilities to address additional EEG features, further enhancing its utility in clinical practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Disclosures

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

We would like to acknowledge the contribution from and thank the Human Epilepsy Project (HEP) experts for gathering and scoring the HEP dataset: Dr. Manu Hegde from the University of California, San Francisco School of

Medicine; Dr. Joon-Yi Kang from Harvard Medical School and Beth Israel Deaconess Medical Center Department of Medicine; Dr. Michael Gelfand from Penn Medicine Department of Neurology; Dr. Liu Lin Thio from St. Louis Children's Hospital Department of Neurology; Drs. Ekrem Kutluay and Zeke Campbell from the Medical University of South Carolina Department. of Neurology and Ralph H. Johnson VA Medical Center; Dr. Ezequiel Gleichgerrcht from Emory University Hospital; Dr. Elizabeth Waterhouse from Richmond VA Medical Center; Dr. Maria Lopez from Bruce W. Carter Department of Veterans Affairs Medical Center; and Dr. Stephan Eisenschenk from the University of Florida and Malcom Randall Department of Veterans Affairs Medical Center.

Dr. Westover's laboratory received support from grants from the National Institutes of Health (NIH) (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598) and the U.S. National Science Foundation (2014431). Dr. Goldenholz was funded by the NIH (K23NS124656). Dr. Taraschenko received salary and research support from the NIH (P20GM130447). Dr. Haider received funding from the National Institute of Neurological Disorders and Stroke (R21NS137117) and the National Center for Advancing Translational Sciences (UG3TR004501-01A1). Dr. Hong was supported by the National Natural Science Foundation of China (62102008). Dr. Halford was supported by the Veterans Affairs Office of Research and Development (I01HX003107-01A2).

References

1. Tatum WO, Rubboli G, Kaplan PW, et al. Clinical utility of EEG in diagnosing and monitoring epilepsy in adults. Clin Neurophysiol 2018;129:1056–1082. DOI: 10.1016/j.clinph.2018.01.019. [PubMed: 29483017]
2. Van Donselaar CA, Schimsheimer RJ, Geerts AT, Declerck AC. Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. Arch Neurol 1992;49:231–237. DOI: 10.1001/archneur.1992.00530270045017. [PubMed: 1536624]
3. Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. Lancet 2019;393:689–701. DOI: 10.1016/S0140-6736(18)32596-0 [PubMed: 30686584]
4. Kane N, Acharya J, Beniczky S, et al. A revised glossary of terms most commonly used by clinical electroencephalographers and updated proposal for the report format of the EEG findings. Revision 2017. Clin Neurophysiol Pract 2017;2:170–185. DOI: 10.1016/j.cnp.2017.07.002. [PubMed: 30214992]
5. Nascimento FA, Barfuss JD, Jaffe A, Westover MB, Jing J. A quantitative approach to evaluating interictal epileptiform discharges based on interpretable quantitative criteria. Clin Neurophysiol 2022;146:10–17. DOI: 10.1016/j.clinph.2022.10.018. [PubMed: 36473334]
6. Jing J, Herlopian A, Karakis I, et al. Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. JAMA Neurol 2020;77:49. DOI: 10.1001/jamaneurol.2019.3531. [PubMed: 31633742]
7. Fiest KM, Sauro KM, Wiebe S, et al. Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. Neurology 2017;88:296–303. DOI: 10.1212/WNL.0000000000003509. [PubMed: 27986877]
8. Kwon C, Wagner RG, Carpio A, Jetté N, Newton CR, Thurman DJ. The worldwide epilepsy treatment gap: a systematic review and recommendations for revised definitions — a report from the ILAE epidemiology commission. Epilepsia 2022;63:551–564. DOI: 10.1111/epi.17112. [PubMed: 35001365]
9. Nascimento FA, Gavvala JR. Education research: neurology resident EEG education: a survey of US neurology residency program directors. Neurology 2016;96:821–824. DOI: 10.1212/WNL.0000000000011354.
10. Benbadis SR, Lin K. Errors in EEG interpretation and misdiagnosis of epilepsy. Eur Neurol 2009;59:267–271. DOI: 10.1159/000115641.
11. Benbadis SR. Errors in EEGs and the misdiagnosis of epilepsy: importance, causes, consequences, and proposed remedies. Epilepsy Behav 2007;11:257–262. DOI: 10.1016/j.yebeh.2007.05.013. [PubMed: 17719853]
12. Kondziella D, Friberg CK, Wellwood I, Reiffurth C, Fabricius M, Dreier JP. Continuous EEG monitoring in aneurysmal subarachnoid hemorrhage: a systematic review. Neurocrit Care 2015;22:450–461. DOI: 10.1007/s12028-014-0068-7. [PubMed: 25277888]

13. Pavel AM, Rennie JM, De Vries LS, et al. A machine-learning algorithm for neonatal seizure recognition: a multicenter randomized controlled trial. Lancet Child Adolesc Health 2020;4:740–749. DOI: 10.1016/S2352-4642(20)30239-X. [PubMed: 32861271]
14. Lam AD, Deck G, Goldman A, Eskandar EN, Noebels J, Cole AJ. Silent hippocampal seizures and spikes identified by foramen ovale electrodes in Alzheimer's disease. Nat Med 2017;23:678–680. DOI: 10.1038/nm.4330. [PubMed: 28459436]
15. Ng MC, Gillis K. The state of everyday quantitative EEG use in Canada: a national technologist survey. Seizure 2017;49:5–7. DOI: 10.1016/j.seizure.2017.05.003. [PubMed: 28501751]
16. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28:31–38. DOI: 10.1038/s41591-021-01614-0. [PubMed: 35058619]
17. Jing J, Dauwels J, Rakthanmanon T, Keogh E, Cash SS, Westover MB. Rapid annotation of interictal epileptiform discharges via template matching under dynamic time warping. J Neurosci Methods 2016;274:179–190. DOI: 10.1016/j.jneumeth.2016.02.025. [PubMed: 26944098]
18. Scheuer ML, Bagic A, Wilson SB. Spike detection: inter-reader agreement and a statistical Turing test on a large data set. Clin Neurophysiol 2017;128:243–250. DOI: 10.1016/j.clinph.2016.11.005. [PubMed: 27913148]
19. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia 2019;60:2037–2047. DOI: 10.1111/epi.16333. [PubMed: 31478577]
20. Jing J, Sun H, Kim JA, et al. Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. JAMA Neurol 2020;77:103. DOI: 10.1001/jamaneurol.2019.3485. [PubMed: 31633740]
21. Tveit J, Aurlien H, Plis S, et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. JAMA Neurol 2023;80:805. DOI: 10.1001/jamaneurol.2023.1645. [PubMed: 37338864]
22. Westover MB, Halford JJ, Bianchi MT. What it should mean for an algorithm to pass a statistical Turing test for detection of epileptiform discharges. Clin Neurophysiol 2017;128:1406–1407. DOI: 10.1016/j.clinph.2017.02.026. [PubMed: 28495216]
23. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE, 2016:761–769. (https://ieeexplore.ieee.org/document/7780458).
24. 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. [PubMed: 36878708]
25. Zhou S, Geng S, Li J, et al. Less is more: reducing overfitting in deep learning for EEG classification. In: 2023 Computing in Cardiology (CinC), 50. Atlanta, GA: IEEE, 2023:1–4. (https://ieeexplore.ieee.org/document/10363878?signout=success).
26. Hong S, Xu Y, Khare A, et al. HOLMES: health online model ensemble serving for deep learning models in intensive care units. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY: Association for Computing Machinery, 2020:1614–1624. (https://dl.acm.org/doi/10.1145/3394486.3403212).
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE, 2016:770–778. (https://ieeexplore.ieee.org/document/7780459).
28. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017:2980–2988.
29. Li M, Wu L, Wiliem A, Zhao K, Zhang T, Lovell B. Deep instance-level hard negative mining model for histopathology images. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI 2019). Shenzhen, China: Springer, 2019:514–522. (https://arxiv.org/abs/1906.09681).
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36. DOI: 10.1148/radiology.143.1.7063747. [PubMed: 7063747]

31. Fürbass F, Kural MA, Gritsch G, Hartmann M, Kluge T, Beniczky S. An artificial intelligence-based EEG algorithm for detection of epileptiform EEG discharges: validation against the diagnostic gold standard. Clin Neurophysiol 2020;131:1174–1179. DOI:10.1016/j.clinph.2020.02.032. [PubMed: 32299000]

32. Fouad A, Azizollahi H, Le Douget JE, et al. Interictal epileptiform discharges show distinct spatiotemporal and morphological patterns across wake and sleep. Brain Commun 2022;4:fcac183. DOI: 10.1093/braincomms/fcac183. [PubMed: 36483575]

33. Karakis I, Pathmanathan JS, Chang R, Cook EF, Cash SS, Cole AJ. Prognostic value of EEG asymmetries for development of drug-resistance in drug-naïve patients with genetic generalized epilepsies. Clin Neurophysiol 2014;125:263–269. DOI: 10.1016/j.clinph.2013.07.028. [PubMed: 24095154]

34. Chinappen DM, Xiao G, Jing J, et al. Spike height improves prediction of future seizure risk. Clin Neurophysiol 2023;150:49–55. DOI: 10.1016/j.clinph.2023.02.180. [PubMed: 37002980]

35. Kim JA, Zheng W-L, Elmer J, et al. High epileptiform discharge burden predicts delayed cerebral ischemia after subarachnoid hemorrhage. Clin Neurophysiol 2022;141:139–146. DOI: 10.1016/j.clinph.2021.01.022. [PubMed: 33812771]

36. Nascimento FA, Jing J, Beniczky S, et al. One EEG, one read — a manifesto towards reducing interrater variability among experts. Clin Neurophysiol 2022;133:68. DOI: 10.1016/j.clinph.2021.10.007. [PubMed: 34814017]

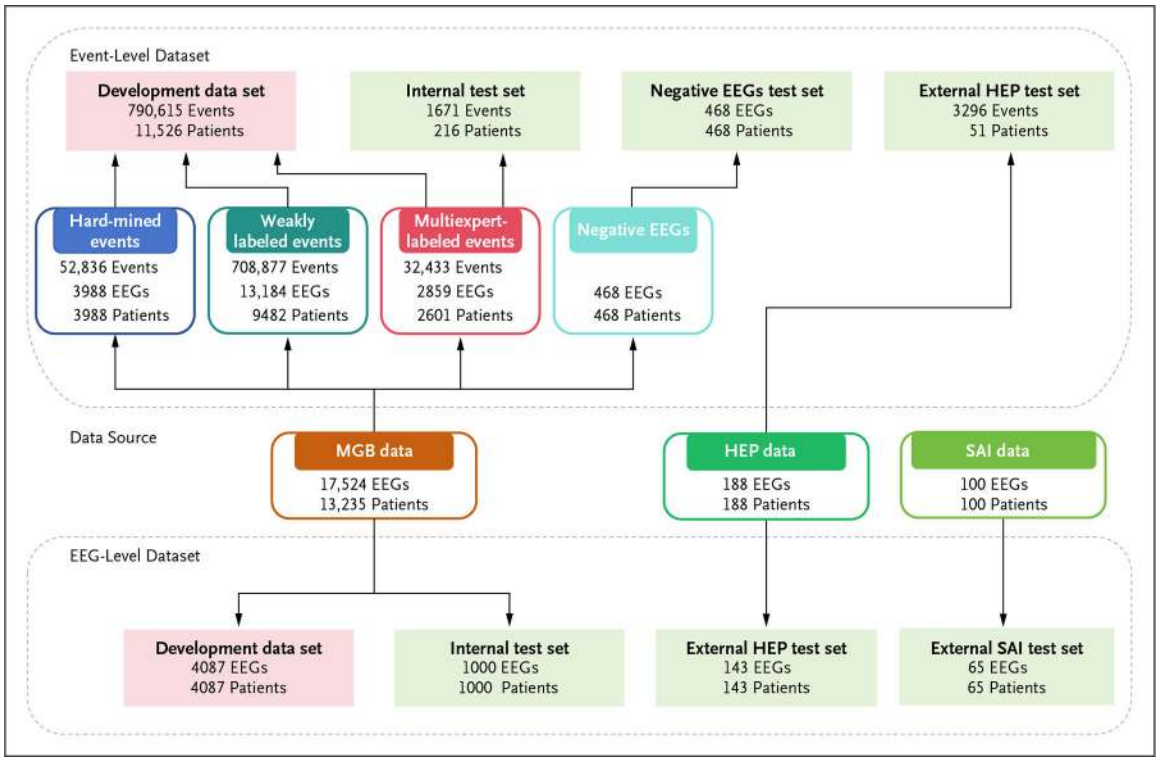

Figure 1. Data Used in Model Development and Validation.

Note that patients and electroencephalograms may overlap in different training phases during model development, but there was strictly no intersection between training and test sets. EEG denotes electroencephalogram; HEP, Human Epilepsy Project; MGB, Massachusetts General Brigham; and SAI, SCORE–Artificial Intelligence.

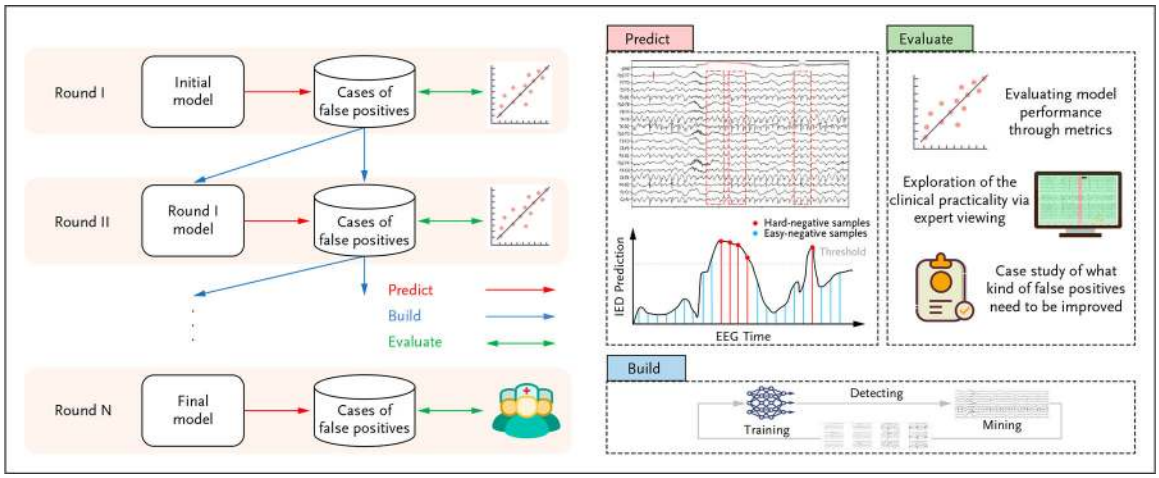

Figure 2. The Pipeline of Hard-Negative Mining.

EEG denotes electroencephalogram; and IED, interictal epileptiform discharges.

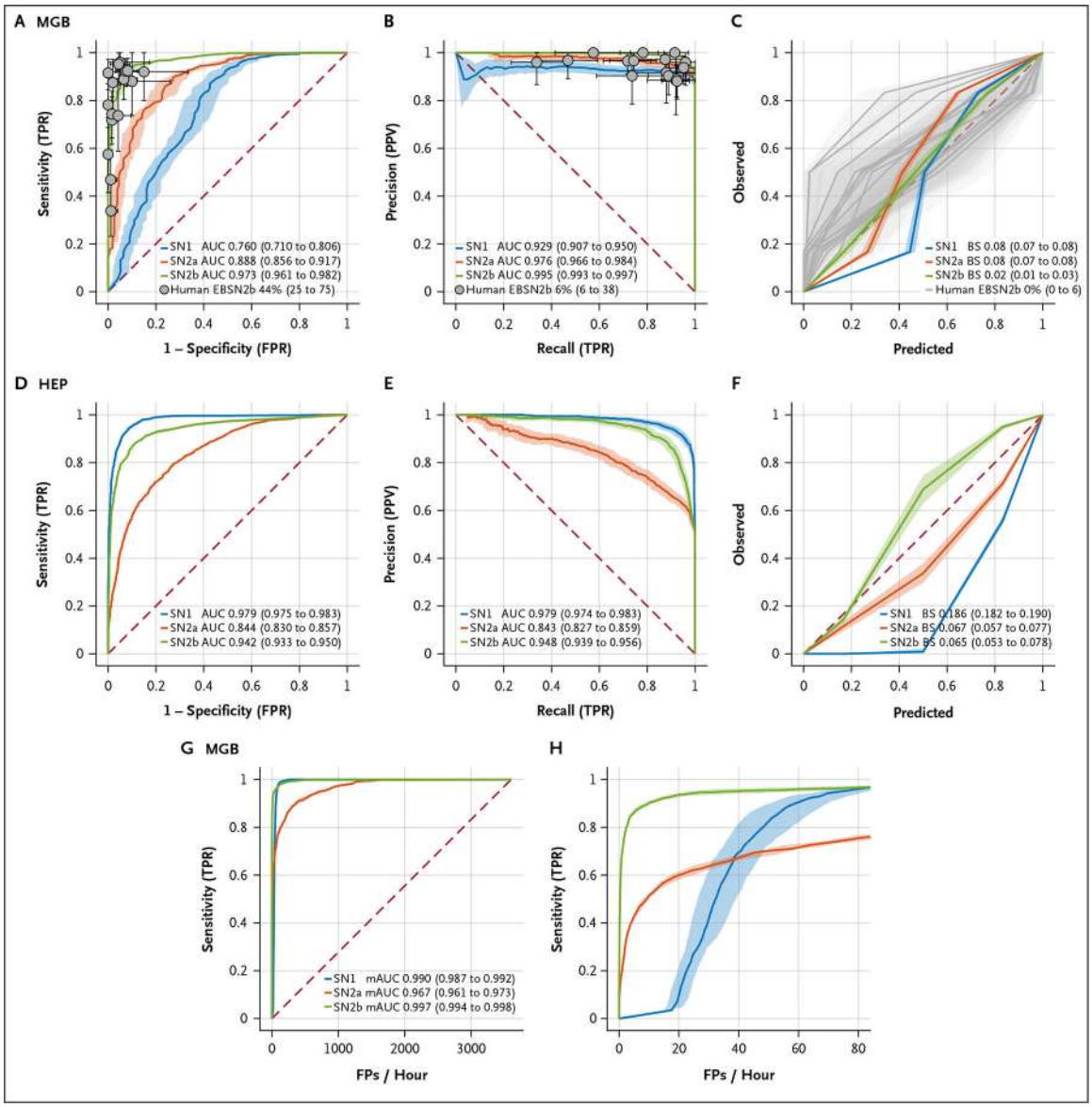

Figure 3. Event-Level Spike-Classification Performance of SpikeNet2 Compared with Benchmark Models.

Panel A shows the receiver operating characteristic (ROC) curve, Panel B the precision–recall (PR) curve, and Panel C the calibration curve for the Massachusetts General Brigham (MGB) test dataset, with 16 human raters' operating points shown for comparison. SpikeNet2 (SN2b) performance is color-coded in green, SpikeNet1 (SN1) in blue, and SpikeNet2 before hard-negative mining (SN2a) in red for comparison. Panel D shows the ROC curve, Panel E the PR curve, and Panel F the calibration curve for SpikeNet2 and comparators on the Human Epilepsy Project external validation dataset. Panel G shows a modified ROC curve and Panel H a zoomed-in modified ROC curve on the MGB control test dataset. Figures in parentheses denote 95% confidence intervals. AUC denotes area under the curve; BS, Brier (calibration) score; EBSN2b, the percentage of experts who outperform SN2b; FP, false positive; FPR, false-positive rate; HEP, Human Epilepsy Project;

mAUC, normalized area under the modified receiver operating characteristic curve; MGB, Massachusetts General Brigham; PPV, positive predictive value; SN1, SpikeNet1; SN2a, SpikeNet2 without hard-negative mining; SN2b, SpikeNet2 with hard-negative mining; and TPR, true-positive rate.

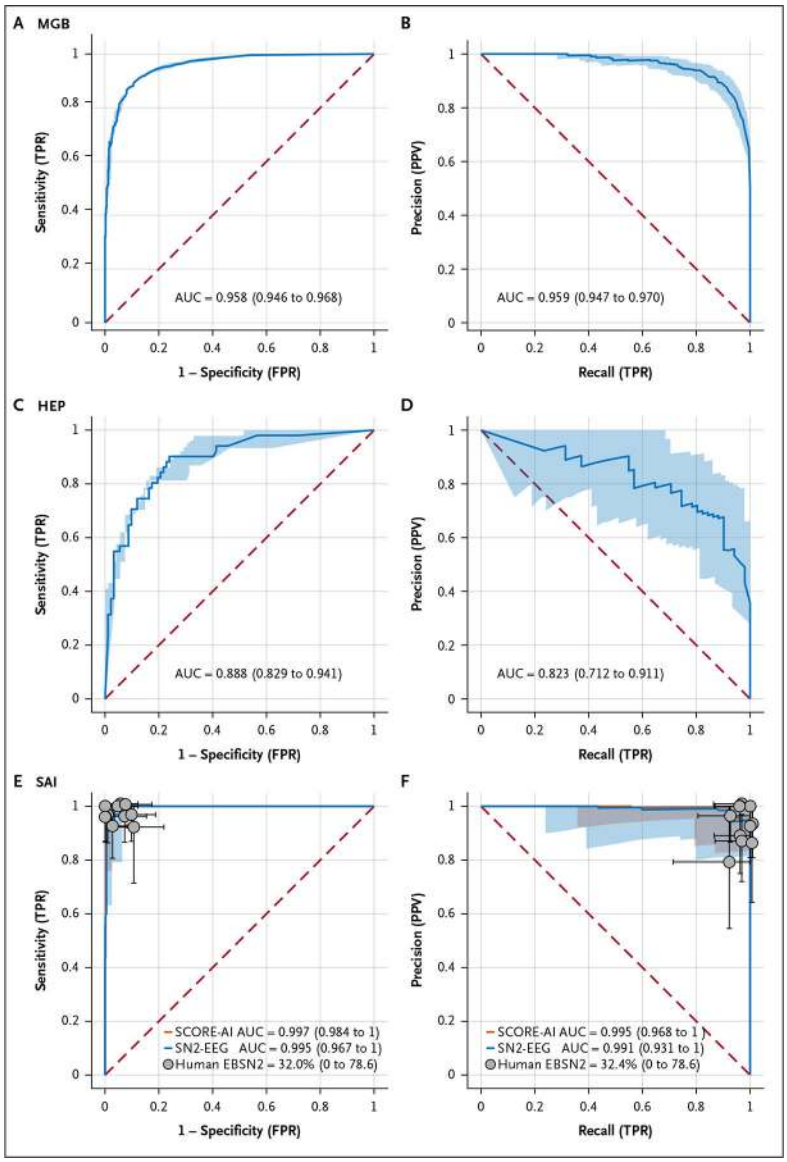

Figure 4. EEG-Level Spike-Classification Performance of SpikeNet2 Compared with Benchmark Models.

Panel A shows the receiver operating characteristic (ROC) curve and Panel B the precision–recall (PR) curve of SpikeNet2 on the Massachusetts General Brigham test set. Panel C shows the ROC curve, and Panel D shows the PR curve of SpikeNet2 on the Human Epilepsy Project external validation dataset. Panel E shows the ROC curve, and Panel F shows the PR curve of SpikeNet2 and the comparator model (SCORE-AI) on the SCORE-AI external validation dataset, with operating points of 14 human raters shown for comparison. Figures in parentheses denote 95% confidence intervals. AUC denotes area under the curve; EBSN2, the percentage of experts who outperform SN2-EEG; FPR, false-positive rate; HEP, Human Epilepsy Project; MGB, Massachusetts General Brigham; PPV, positive predictive value; SAI, SCORE–Artificial Intelligence; SN2-EEG, SpikeNet2 for electroencephalography-level task; and TPR, true-positive rate.

Table 1.

Cohort Characteristics of the Internal Massachusetts General Brigham Dataset.*

Characteristic	N (%)
Age group — no. of patients (%)	
0–1 month	95 (0.72)
1–3 months	51 (0.39)
3–6 months	53 (0.40)
6–12 months	72 (0.54%)
1–10 years	1031 (7.79)
10–20 years	1160 (8.76)
20–30 years	1296 (9.79)
30–40 years	1085 (8.20)
40–50 years	1176 (8.89)
50–60 years	1762 (13.31)
60–70 years	2121 (16.03)
70–80 years	1774 (13.40)
80–90 years	1027 (7.76)
90–100 years	205 (1.55)
≥100 years	6 (0.05)
Unknown	321 (2.43)
Age — median (IQR)	53 (43) years
Sex — no. of patients (%)	
Female	6242 (47)
Male	6672 (50.4)
Unknown	321 (2.4)
Race — no. of patients (%)†	
White	9556 (72)
Black/African American	805 (6)
Asian	441 (3)
Multiracial	117 (1)
Other races‡	786 (6)
Unknown	1530 (12)
Neurodevelopmental state — no. of patients (%)	
Autism	671 (5)
Developmental delay	1318 (10)
Other states	11,113 (84)
Unknown	558 (4)
EEG setting — no. of EEGs (%)	
Routine§	10,202 (58)
EMU	1032 (6)
LTM	5977 (34)
Unknown	313 (2)

[*] Demographic information is unavailable for the external Human Epilepsy Project dataset. EEG denotes electroencephalogram; EMU, epilepsy monitoring unit; IQR, interquartile range; and LTM, long-term monitoring unit or intensive care unit.

[†] Race obtained via patient database.

[‡] "Other races" include minority groups such as American Indian or Alaska Native, Native Hawaiian, Other Pacific Islander, and individuals originally recorded as "Other race" in the database, for whom we have limited data access.

[§] Routine setting refers to the outpatient clinic.

Table 2.

Cohort Characteristics of the External SCORE-AI Dataset.*

	Pediatric (<16 Years), N		Adult (≥16 Years), N		
EEG Country of Origin	**Normal**	**Abnormal**	**Normal**	**Abnormal**	**Total**
United States	2	6	6	11	25
Denmark	4	3	6	7	20
Norway	9	11	16	19	55
Total	15	20	28	37	100

* EEG denotes electroencephalogram.