Received: 10 January 2022 | Revised: 17 February 2022 | Accepted: 17 February 2022

DOI: 10.1111/epi.17206

Epilepsia®

RESEARCH ARTICLE

Accurate identification of EEG recordings with interictal epileptiform discharges using a hybrid approach: Artificial intelligence supervised by human experts

Mustafa Aykut Kural[1,2,3] | **Jin Jing**[4] | **Franz Fürbass**[5] | **Hannes Perko**[5] | **Erisela Qerama**[2,3] | **Birger Johnsen**[2,3] | **Steffen Fuchs**[2] | **M. Brandon Westover**[4] | **Sándor Beniczky**[1,2,3]

[1]Department of Clinical Neurophysiology, Danish Epilepsy Centre Filadelfia, Dianalund, Denmark

[2]Department of Clinical Neurophysiology, Aarhus University Hospital, Aarhus, Denmark

[3]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

[4]Department of Neurology, Harvard Medical School, Massachusetts General Hospital, Boston, Massachusetts, USA

[5]Center for Health & Bioresources, AIT Austrian Institute of Technology GmbH, Vienna, Austria

Correspondence
Sándor Beniczky, Aarhus University Hospital and Danish Epilepsy Center, Visby Allé 5, 4293, Dianalund, Denmark.
Email: sbz@filadelfia.dk

Funding information
Foundation for the National Institutes of Health, Grant/Award Number: 1R01NS107291

Abstract

Objective: To evaluate the diagnostic performance of artificial intelligence (AI)–based algorithms for identifying the presence of interictal epileptiform discharges (IEDs) in routine (20-min) electroencephalography (EEG) recordings.

Methods: We evaluated two approaches: a fully automated one and a hybrid approach, where three human raters applied an operational IED definition to assess the automated detections grouped into clusters by the algorithms. We used three previously developed AI algorithms: Encevis, SpikeNet, and Persyst. The diagnostic gold standard (epilepsy or not) was derived from video-EEG recordings of patients' habitual clinical episodes. We compared the algorithms with the gold standard at the recording level (epileptic or not). The independent validation data set (not used for training) consisted of 20-min EEG recordings containing sharp transients (epileptiform or not) from 60 patients: 30 with epilepsy (with a total of 340 IEDs) and 30 with nonepileptic paroxysmal events. We compared sensitivity, specificity, overall accuracy, and the review time-burden of the fully automated and hybrid approaches, with the conventional visual assessment of the whole recordings, based solely on unrestricted expert opinion.

Results: For all three AI algorithms, the specificity of the fully automated approach was too low for clinical implementation (16.67%; 63.33%; 3.33%), despite the high sensitivity (96.67%; 66.67%; 100.00%). Using the hybrid approach significantly increased the specificity (93.33%; 96.67%; 96.67%) with good sensitivity (93.33%; 56.67%; 76.67%). The overall accuracy of the hybrid methods (93.33%; 76.67%; 86.67%) was similar to the conventional visual assessment of the whole recordings (83.33%; 95% confidence interval [CI]: 71.48–91.70%; $p > .5$), yet the time-burden of review was significantly lower ($p < .001$).

Significance: The hybrid approach, where human raters apply the operational IED criteria to automated detections of AI-based algorithms, has high specificity,

good sensitivity, and overall accuracy similar to conventional EEG reading, with a significantly lower time-burden. The hybrid approach is accurate and suitable for clinical implementation.

KEYWORDS
artificial intelligence, automated detection, convolutional neural networks, deep learning, hybrid method, interictal epileptiform EEG discharges

1 | INTRODUCTION

Interictal epileptiform discharges (IEDs) are well-documented electrophysiological biomarkers for epilepsy and they are broadly used in clinical practice.[1] In skilled hands, IEDs provide valuable information in the diagnostic workup of patients suspected for seizures and epilepsy.[2] However, this expertise is scarce, and the increasing number and duration of EEGs recorded worldwide have led to an increased workload.[1] In the past 45 years, many attempts at automated IED detection have been published.[3] Despite the significant development, automated detection of IED has not yet been widely implemented in clinical electroencephalography (EEG) reading, and compelling evidence for their accuracy and performance is still lacking.[3]

Artificial intelligence (AI) is increasingly recognized as a useful tool in health care applications, including epilepsy.[4] With the development of deep learning methods, interest in using AI algorithms for IED detection is growing.[3] An alternative approach to fully automated detection of IEDs is using hybrid (semi-automated) systems, in which human raters evaluate the IED candidate waves, automatically detected and clustered (grouped) by the algorithm into IED types.[5] This approach potentially combines the high sensitivity of automated detection with the high specificity from the human experts, and is expected to significantly decrease the workload.

Most of the previously published validation studies were affected by numerous sources of bias: very small data sets, lack of control and of distractor data, and using the same data set for training and testing.[3] Even well-conducted studies avoiding these flaws bear important potential sources of error, due to a lack of an unequivocal external gold standard, and evaluating short EEG segments instead of the whole recording. Using expert opinion as gold standard for single IEDs in short EEG segments is questionable, because inter-rater agreement is low,[6] it is uncertain that the majority decision is correct,[7] and when most raters are trained in the same school, a systemic (institutional) bias may become inherent in the data set. In addition, exaggeratedly optimistic estimates of specificity may occur when short EEG segments are evaluated instead of the whole EEG recording.

Key points

- We evaluated the diagnostic performance of artificial intelligence (AI) algorithms for identifying the presence of interictal epileptiform discharges in routine electroencephalography (EEG) recordings
- We evaluated two approaches: a fully automated one and a hybrid approach, where human raters assessed the automated detections
- The fully automated approach had high sensitivity but low specificity
- The hybrid approach had high specificity, good sensitivity, and accuracy similar to that of conventional assessment
- The time-burden of review was significantly lower for the hybrid approach compared with the conventional reading of the whole recording

In an EEG recording some IEDs appear well defined, whereas others are distorted to different degrees by superimposed background activity and lower amplitude of the IEDs.[8] Evidence for these "hidden" IEDs has been provided from studies analyzing topographic voltage maps, compared with IED-related hemodynamic changes in EEG-fMRI (functional magnetic resonance imaging), and from comparison with intracranial recordings.[9–11] Hence, there is a considerable gray zone between IEDs and nonepileptiform sharp transients, and the question "is this discharge an IED" is ill-posed from a clinical point of view. A more meaningful clinical question is rather "does this EEG recording include IEDs?"[12] In other words: does the EEG recording have the biomarker for epilepsy?

Cutoff values (decision thresholds) are adjustable in IED-detection algorithms. Although receiver-operating characteristic (ROC) curves are useful to determine the optimal cutoff values in training data sets, validation studies must use pre-defined cutoff values.[13] In the validation data set, choosing the cutoff value post hoc (after the analysis and comparison with the gold standard) results in overfitting and overoptimistic evaluation of the

algorithms. Therefore, in this study we used predefined cutoff values for validating the previously developed algorithms.

In this study we present an external validation of fully automated and semi-automated (hybrid) IED detection. We performed head-to-head comparison of three previously developed AI-based algorithms. Two of them were using convolutional neural networks: SpikeNet[14] and the commercially available software package Encevis using the DeepSpike algorithm for IED detection.[15] The spike detector in the commercially available software program Persyst uses extracted features and a feed-forward neural network.[16] In the hybrid approach, clusters of IED candidate waves, detected by the algorithms, were visually evaluated by experts using the criteria in the operational definition of the International Federation of Clinical Neurophysiology (IFCN) for IEDs,[17] which were shown previously to provide high specificity, essential in clinical EEG reading.[18] To circumvent the shortcomings and potential bias described above, we used an external diagnostic gold standard, derived from assessment of the patients´ habitual clinical episodes (epileptic or not), and we evaluated continuous 20-min recordings, corresponding to routine EEG.[1] We determined diagnostic accuracy (sensitivity and specificity) at the recording level (EEG including IEDs or not), and calculated the change in time consumption of EEG reading.

2 | METHODS

2.1 | Patients and EEG recordings

Inclusion criteria were: patients 1 year or older, who had their diagnosis (epilepsy or non-epileptic paroxysmal events) based on video-EEG recordings of their habitual clinical episode, and who had sharp transients (epileptiform or not) recorded during a 20-min EEG—corresponding to a routine EEG recording.[1] Sharp transients were defined as having pointed peak and amplitude higher than the background activity. These were marked in the 20-min EEG recordings by two of the authors (MAK and SB) who did not participate in the subsequent, blinded rating of the EEG recordings. Patients with epilepsy had an additional inclusion criterion: the marked sharp transients in their routine (20-min) EEG recording had to be concordant with the ictal event in the gold standard (ie, bilateral synchronous IEDs for generalized seizures in patients with generalized epilepsies, and focal with peak negativity located in the same sub-lobar region as the ictal recordings acquired in patients with focal onset seizures). Patients with inconclusive video-EEG recordings and patients having both epileptic and nonepileptic seizures were excluded, because we needed unequivocal diagnostic gold standard for each patient.

We aimed for an equal number of patients with and without epilepsy, because the analysis outcome was a dichotomous classification (epilepsy or not). The distractors (nonepilepsy patients) had paroxysmal nonepileptic events, and to challenge (stress test) the algorithm, their EEG recordings included nonepileptiform sharp transients (such as: spiky fluctuations of the background, wicket waves, small sharp spikes, and artifacts). We analyzed 60 EEG recordings (of 20 min each) from 30 consecutive patients with epilepsy and 30 consecutive patients with nonepileptic paroxysmal events who fulfilled the inclusion criteria. The median age of the 60 patients (33 female) was 36 years (range: 2–77 years). The group of nonepileptic patients included psychogenic nonepileptic seizures (12 patients), sleep disorders (10 patients), paroxysmal movement disorders (4 patients), and syncope (4 patients). A total of 340 IEDs were marked by the experts who did not participate in the subsequent, blinded rating of the EEG studies (MAK and SB). Three hundred six IEDs were focal (recorded in 26 patients with temporal lobe and 2 patients with extratemporal lobe epilepsy), and 34 IEDs were generalized (recorded in 2 patients). The nonepileptic sharp transients recorded in 30 patients with nonepileptic paroxysmal events included: spiky fluctuation of the background/wicket spikes ($n = 58$), spiky artifacts ($n = 12$), small sharp spikes ($n = 5$), spiky vertex sharp-waves ($n = 2$), and small sharp transients intermixed with hypnagogic hypersynchrony ($n = 2$). EEG was recorded at the Danish Epilepsy Center and Aarhus University Hospital (Denmark), using the IFCN array of 25 electrodes, including the inferior temporal electrode chain,[19] using NicoletOne EEG equipment (Natus Neuro), at a sampling frequency of 0.5 kHz. The regional ethics committee reviewed the protocol. Written informed consent from the patients was not needed for this non-interventional study, using retrospective analyses of de-identified data.

2.2 | Diagnostic gold standard and index tests

To avoid circular thinking and the uncertainty of identifying IEDs (see Introduction), the diagnostic gold standard was external, that is, based on a modality different from IEDs: video-EEG recordings of the patient´s habitual clinical events (epileptic seizure or paroxysmal nonepileptic events), which is the most reliable diagnostic method in these patients.[1] Ictal EEG and semiology were evaluated as part of the patients' diagnostic workup, and then re-evaluated by two of the authors (MAK and SB) for this

study. Classification (epilepsy or not) was based on consensus discussions of the authors.

In diagnostic accuracy studies,[20] the term *index test* refers to the methods evaluated in the study, which are compared with the diagnostic gold standard (reference standard). In this study, the index tests were: conventional reading of routine (interictal, 20 min) EEG, fully automated IED detection, and the hybrid approach.

2.3 | Conventional EEG reading

Three expert raters (BJ, EQ, and SF), board certified in clinical neurophysiology, with more than 10 years of experience in EEG reading, independently evaluated the 20-min EEG recordings, blinded to all other data. They visually evaluated the whole EEG recordings and marked IEDs solely based on their expert opinion. Time for the visual assessment was recorded for each EEG and each rater. In this article, we refer to this as the conventional EEG reading. Majority consensus scorings were used for further analysis.

2.4 | Fully automated IED detection and the hybrid method

We used three different, previously developed, AI-based IED-detection algorithms. Encevis is a Conformité Européenne (CE) marked and Food and Drug Administration (FDA) cleared EEG software package that uses the DeepSpike algorithm for detection of IEDs. DeepSpike was developed using the Fast Region-based Convolutional Network method (Fast R-CNN).[15] It uses deep regression for estimating the position of EDs (negative peaks)[21,22] followed by classification of EDs.[15] Supervised and unsupervised learning was used to train DeepSpike. For supervised learning, 447 000 labeled EEG epochs from 166 patients and synthetic data sets were used.[15] For unsupervised learning, 590 000 epochs from 289 patients of the publicly available EEG Corpus of the Temple University were used.[23] The cutoff value for detection was 0.3. Detections were clustered using the maximum electrode position and EEG waveforms were presented as overlay plots in Encevis (Appendix S1). SpikeNet[14] was developed using a convolutional neural network, with an architecture similar to that in Hannun et al.[24] It was trained in two steps, using 9571 scalp EEG records with and without IEDs.[14]The detection threshold was a score of 0.4. For this project, we augmented SpikeNet with two new features. First, a background rejection method was developed specifically for this project, and its threshold parameter was set to $\theta_b = 50$. Second, a clustering method was developed wherein detected IEDs were clustered according to the leading channel (Appendix S2). Persyst P13 spike detector is commercially available (Persyst). It was developed using 20 feedforward neural network rules to describe the morphology, field, and context of each event.[16,25] Artifact reduction was activated and the cutoff perception value was set to 0.5 (medium sensitivity) in this study. Detected events were clustered according to the electrode with the peak negativity.

In the hybrid method, the three expert raters (BJ, EQ, and SF) independently reviewed the automated detections, with IED candidate waves clustered into IED types by the algorithms, as described above and in Appendixes S1 and S2. They were instructed to implement the six criteria of the operational definition of IEDs (Kane et al. 2017),[17] and decide whether the automated detections in each cluster were IEDs or not. They did not review the whole EEG, but only the detected and clustered waveforms. The raters were allowed to change montages (bipolar and common average), digital filters, and gain, when reviewing the EEG recordings. Voltage maps and the inferior temporal electrodes were available in Encevis and Persyst, but not in SpikeNet. Time for the visual assessment was recorded for each EEG and each rater and algorithm. The raters evaluated detections of each algorithm in different, randomized order, with at least 2 months between reading sessions. Majority consensus decisions for each recording and each algorithm were considered for further analysis.

2.5 | Evaluation of the analysis results and statistics

To circumvent the shortcomings related to the lack of a gold standard for individual IEDs, we evaluated the accuracy at the EEG recording level. For the fully automated method, EEG recordings were labeled positive if they had automated detections and negative if they did not. For the hybrid method, EEG recordings were labeled positive if they had automated detections confirmed by the expert raters (BJ, EQ, and SF) and negative if they did not. Then we compared them with the diagnostic gold standard: recordings labeled positive were considered true positives (TPs) if the patients had epilepsy and the detections coincided with the markings of the experts who did not rate the index tests (MAK and SB) or had the same voltage topography. They were considered false positives (FPs) if the patients had nonepileptic paroxysmal events or if the detections were not concordant with the expert markings (either coincidence in time or with the same voltage distribution) in a patient with epilepsy. Recordings labeled negative were considered true negatives (TNs) if the patients had

paroxysmal nonepileptic events, and false negatives (FNs) if they had epilepsy. We used the conventional formulas for the diagnostic accuracy measures: sensitivity = TP/(TP + FN), specificity = TN/(TN + FP), accuracy = (TP + TN)/(TP + TN + FP + FN), and we calculated area under the ROC (AUROC) curve from the scorings of the three blinded raters.[26] Because EEG over-reading (ie, over-interpretation of nonepileptiform sharp transients) is the most common cause of misdiagnosing epilepsy, our goal was to achieve a high specificity (over 90%).[27–29]

We calculated 95% confidence intervals (CIs) for the diagnostic accuracy measures using Wilson´s method.[30] We compared the diagnostic accuracy measures of the different analysis methods using McNemar's test.[31] After analyzing the data distribution using the Kolmogorov-Smirnov test for normality, we used *t* tests for dependent samples to compare the time consumption of conventional EEG reading with the hybrid methods.[32]

3 | RESULTS

Figure 1 shows examples with the automated detections, grouped into clusters by the three software packages. Table 1 summarizes the diagnostic accuracy measures of the fully automated analysis and the hybrid approach (applying the IFCN IED criteria on the automated detections), compared with the conventional EEG reading (expert opinion based on visual analysis of the whole recording, unrestricted by any criteria).

For all three AI algorithms, specificity of the fully automated approach was too low for clinical implementation—only SpikeNet approached a reasonable specificity, although at the cost of lower sensitivity (Table 1 and Figure 2). Specificity increased significantly for all three algorithms when using the hybrid approach, reaching a high level (93%–97%) for all three algorithms (Table 1 and Figure 2). Sensitivity of the fully automated approach was high (67%–100%) for all three algorithms (Table 1). Although this decreased when using the hybrid approach, the decrease was significant only for Persyst, which had a 100% sensitivity with the fully automated approach (Table 1 and Figure 2).

The accuracy of the hybrid approaches was between 77% and 93% (Table 1), with AUROC between 0.76 and 0.93 (Encevis = 0.935; Persyst = 0.871; SpikeNet = 0.757), which is similar to the conventional evaluation of the EEG recordings by experts (accuracy: 83%; AUROC: 0.837), and suitable for clinical implementation (Appendixes S3 and S4). The mean time consumption for conventional EEG assessment was 160 s per recording (95% CI: 146–172 s). This was reduced by 26%–91% using the hybrid approach: 14 s (95% CI: 11–16 s) with SpikeNet, 59 s (95% CI: 50–67 s) with Persyst, and 118 s (95% CI: 103–134 s) with Encevis. Time burden was significantly shorter for the hybrid approach (in all three software packages) as compared with the conventional reviewing of the entire EEG ($p < .001$, for all three software packages).

4 | DISCUSSION

In current practice, clinical EEG reading is based on visual assessment of the whole EEG recordings.[1] However, this is time-consuming and requires extensive training to classify EEG patterns correctly.[33] We have demonstrated that hybrid systems combine high sensitivity from the AI-based algorithms with high specificity from human raters, and significantly reduce the time-burden of reviewing the EEG recording. Humans do not need to browse the whole recording: they inspect only the IED-candidate waves detected and clustered by the algorithms. Implementing the operational IED definition of the IFCN in this decision-making yields high specificity, surpassing conventional EEG reading. Previous studies showed that trainees could easily learn the IFCN criteria, significantly improving their accuracy and interrater agreement.[34] Hence, using the hybrid approach in clinical practice could help improve the quality of care and decrease the workload.

Although many algorithms for automated IED detection have been developed in the past 45 years, they are rarely used in clinical practice.[3] This is partly due to the discordance between the published performance of the algorithms and what clinicians experience when implementing them in real-world settings: a frustratingly high rate of false detections. Two aspects related to the tradition established in this field may have contributed to the overoptimistic estimation of the algorithms: using expert opinion on IEDs instead of an external gold standard, and focusing on individual candidate IEDs instead of the whole recording.[3] A false detection in an EEG recording erroneously classifies it as including IED. However, when specificity is calculated from the short, 2 s segments, the specificity appears to be high in the misclassified recording (599 segments without detection divided by the total of 600 segments gives a specificity of 99%). To circumvent these shortcomings, we used an external diagnostic gold standard, derived from the video-EEG recordings of the patients' habitual clinical episodes, and we evaluated the performance at the recording level. Because IEDs are used as biomarkers for epilepsy, we asked the clinically relevant question: does this EEG recording include IEDs? IED morphology varies in different age-groups.[35] Our data set included a wide age range (2–77 years).

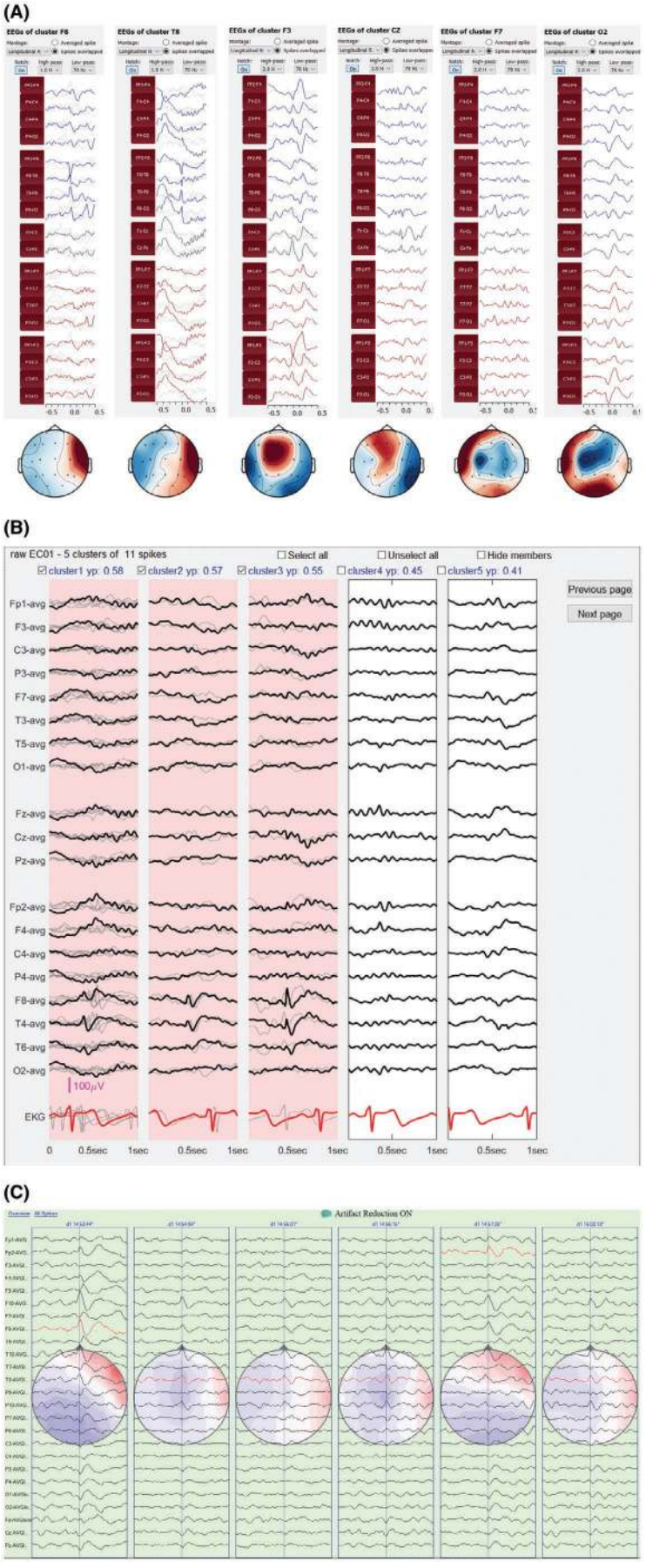

FIGURE 1 Examples with the automated detections, grouped into clusters by the three software packages (A: Encevis; B: SpikeNet; C: Persyst)

TABLE 1 Diagnostic accuracy measures of the fully automated analysis and the hybrid approach, compared with the conventional EEG reading

	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
Conventional visual assessment	93.33% (77.93–99.18)	73.33% (55.11–87.72)	83.33% (71.48–91.70)
Fully automated			
Encevis	96.67% (82.78–99.92)	16.67% (5.64–34.72)	56.67% (43.24–69.41)
SpikeNet	66.67% (47.19–82.71)	63.33% (43.86–80.07)	65.00% (51.60–76.87)
Persyst	100.00% (88.43–100.00)	3.33% (0.08 –17.22)	51.67% (38.39–64.77)
Hybrid approach			
Encevis	93.33% (77.93–99.18)	93.33% (77.93–99.18)	93.33% (77.93–99.18)
SpikeNet	56.67% (37.43–74.54)	96.67% (82.78–99.92)	76.67% (63.96–86.62)
Persyst	76.67% (57.72–90.07)	96.67% (82.78–99.92)	86.67% (75.41–94.06)

As in the real world, the number of EEG epochs without IEDs in our data set exceeded by far the number with IEDs (340 s with IEDs of 72 000 s EEG recording). Furthermore, to stress test the specificity of the algorithms, all 30 distractor EEG recordings contained nonepileptiform sharp transients from patients with nonepileptic paroxysmal events. A high specificity is essential in clinical EEG reading, because over-reading (over-interpretation of sharp transients) is the most common cause of misdiagnosing epilepsy.[27–29] It is estimated that one third of patients seen at epilepsy centers for drug-resistant seizures do not have epilepsy[36–39] causing many detrimental consequences for the patients, such as restricting career choices, driving, unnecessary anti-seizure medications, and not treating the correct diagnosis.[40,41] Our findings show that despite the high sensitivity, the currently available AI algorithms and predefined cutoff values do not reach a sufficient specificity for clinical applications. However, human experts can rapidly evaluate the automated detections, clustered into IED types, adding a high specificity to this hybrid approach when using the IFCN criteria. Clustering is important especially for longer recordings: the user does not need to review all individual examples within a cluster to classify the cluster, which saves a considerable amount of time in the case of long-term recordings.[5]

Although the head-to-head comparison of the three algorithms showed some difference in the performance of the fully automated detection (some had better sensitivity, others better specificity), there was no clear winner, and the results rather reflect the current state-of-art of AI-based IED detection—in general. Concerning the hybrid approach, all three software packages had similar accuracy, noninferior to the conventional EEG reading by experts. The specificity of the hybrid approach was higher than the traditional, expert-assessment, probably due to the systematic application of the IFCN criteria. The sensitivity of the hybrid approach was similar to that of the expert assessment for two algorithms. The current version of the SpikeNet software does not use and display the inferior temporal electrodes of the standard IFCN array[19] and does not include voltage maps, which in part explain the lower sensitivity of the hybrid method using this algorithm. These aspects could easily be added in an updated version. In addition, the background rejection and clustering methods used for SpikeNet were developed specifically for this work, and could be further optimized.

Approximately one of six patients with focal epilepsy do not have IEDs during long-term video-EEG monitoring.[42] The presence of IEDs was an inclusion criterion for the patients with epilepsy in this study. Therefore, the sensitivity values are higher than what is expected in the general population. According to the inclusion criteria, all patients with nonepileptiform sharp transients had nonepileptic paroxysmal events, and none of them had epileptic seizures during the long-term monitoring. Therefore, specificity was not affected by patients with epilepsy having nonepileptiform sharp transients, but no IEDs during the long-term monitoring.[43]

One major limitation of this study is that the inclusion criterion (patients who undergone long-term video-EEG monitoring) might be too restrictive and the included patients may not be representative of the wide variety of IED morphologies encountered in practice. Because we included consecutive patients admitted to the epilepsy monitoring unit, focal epilepsy (especially temporal lobe epilepsy) is overrepresented in this study. Indeed, we gave higher priority to the reliability of the external diagnostic gold standard than the possible selection bias. The identification of generalized IEDs is less challenging than focal IEDs. A second limitation is the relatively small size of our data set ($N = 60$ patients). There is a need for large, multicenter databases with EEG recordings from patients with epilepsy and distractor conditions (nonepileptic paroxysmal events), with diagnoses validated by long-term video-EEG studies or by long-term follow-up, to further elucidate this.[7]

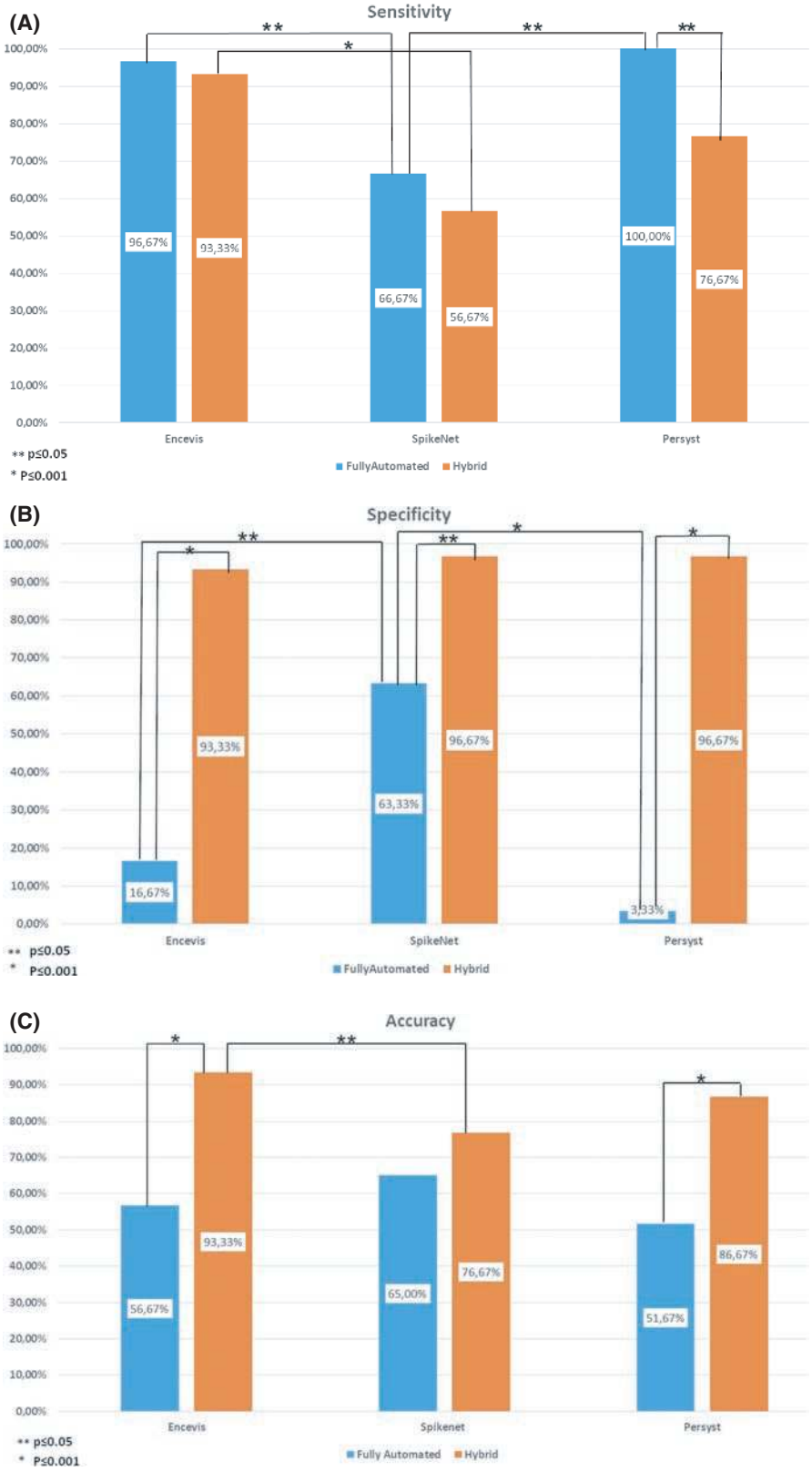

FIGURE 2 Comparisons of the fully automated and hybrid methods, using the three algorithms (A: sensitivity, B: specificity, C: accuracy). Statistically significant differences are marked by asterisks

In conclusion: the hybrid approach where human raters use the operational IFCN definition to confirm IEDs automatically detected and clustered by AI-based algorithm accurately distinguishes routine EEG recordings having IEDs from those that do not. The hybrid method decreases the workload and is suitable for clinical implementation. More research and development is needed before fully automated IED detection can perform as well as the hybrid approach.

ACKNOWLEDGMENT

MBW and JJ receive funding from the National Institutes of Health (NIH; 1R01NS107291). We would like to express our gratitude to Persyst for guidance and advice concerning the use of their spike-detector software.

CONFLICT OF INTEREST

MBW is a co-founder of Beacon Biosignals, which played no role in this work. FF and HP are employees of the AIT Austrian Institute of Technology, which owns all rights to the trademark Encevis and is the manufacturer of the software package Encevis. The remaining authors have no conflict of interest. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

ORCID

Jin Jing https://orcid.org/0000-0002-2415-5854
Franz Fürbass https://orcid.org/0000-0002-6744-2802
Sándor Beniczky https://orcid.org/0000-0002-6035-6581

REFERENCES

1. Tatum WO, Rubboli G, Kaplan PW, Mirsatari SM, Radhakrishnan K, Gloss D, et al. Clinical utility of EEG in diagnosing and monitoring epilepsy in adults. Clin Neurophysiol. 2018;129(5):1056–82.
2. Pillai J, Sperling MR. Interictal EEG and the diagnosis of epilepsy. Epilepsia. 2006;47(suppl 1):14–22.
3. da Silva LC, Tjepkema-Cloostermans MC, van Putten MJAM. Machine learning for detection of interictal epileptiform discharges. Clin Neurophysiol. 2021;132(7):1433–43.
4. Beniczky S, Karoly P, Nurse E, Ryvlin P, Cook M. Machine learning and wearable devices of the future. Epilepsia. 2021;62(Suppl 2):S116–24.
5. Scherg M, Ille N, Weckesser D, Ebert A, Ostendorf A, Boppel T, et al. Fast evaluation of interictal spikes in long-term EEG by hyper-clustering. Epilepsia. 2012;53(7):1196–204.
6. Halford JJ, Arain A, Kalamangalam GP, LaRoche SM, Leonardo B, Basha M, et al. Characteristics of EEG interpreters associated with higher Interrater agreement. J Clin Neurophysiol. 2017;34:168–73.
7. Nascimento FA, Jing J, Beniczky S, Benbadis SR, Gavvala JR, Yacubian EMT, et al. One EEG, one read - A manifesto towards reducing interrater variability among experts. Clin Neurophysiol. 2022;133:68–70.
8. Gotman J, Ives JR, Gloor P. Automatic recognition of inter-ictal epileptic activity in prolonged EEG recordings. Electroencephalogr Clin Neurophysiol. 1979;46(5):510–20.
9. Grouiller F, Thornton RC, Groening K, Spinelli L, Duncan JS, Schaller K, et al. With or without spikes: localization of focal epileptic activity by simultaneous electroencephalography and functional magnetic resonance imaging. Brain. 2011;134(Pt 10):2867–86.
10. Kr V, Rajagopalan SS, Bhardwaj S, Panda R, Reddam VR, Ganne C, et al. Machine learning detects EEG microstate alterations in patients living with temporal lobe epilepsy. Seizure. 2018;61:8–13.
11. Baldini S, Pittau F, Birot G, Rochas V, Tomescu MI, Vulliémoz S, Seeck M. Detection of epileptic activity in presumably normal EEG. Brain Commun. 2020;2(2):fcaa104. https://doi.org/10.1093/braincomms/fcaa104
12. Gotman J, Gloor P, Ray WF. A quantitative comparison of traditional reading of the EEG and interpretation of computer-extracted features in patients with supratentorial brain lesions. Electroencephalogr Clin Neurophysiol. 1975;38(6):623–39.
13. Beniczky S, Ryvlin P. Standards for testing and clinical validation of seizure detection devices. Epilepsia. 2018;59(Suppl 1):9–13.
14. Jing J, Sun H, Kim JA, Herlopian A, Karakis I, Ng M, et al. Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. JAMA Neurol. 2020;77(1):103–8.
15. 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(6):1174–9.
16. 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(1):243–50.
17. Kane N, Acharya J, Beniczky S, Caboclo L, Finnigan S, Kaplan PW, 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;2:170–85.
18. Kural MA, Duez L, Sejer Hansen V, Larsson PG, Rampp S, Schulz R, et al. Criteria for defining interictal epileptiform discharges in EEG: a clinical validation study. Neurology. 2020;94(20):e2139–47.
19. Seeck M, Koessler L, Bast T, Leijten F, Michel C, Baumgartner C, et al. The standardized EEG electrode array of the IFCN. Clin Neurophysiol. 2017;128:2070–7.
20. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
21. Ren S, He K, Girshick R. Faster R-CNN: Towards real-time object detection with region proposal networks. ArXiv. 2015: 150601497 Cs.
22. Belagiannis V, Rupprecht C, Carneiro G, Navab N. Robust optimization for deep regression. ArXiv. 2015: 150506606 Cs.
23. Obeid I, Picone J. The temple university hospital EEG data corpus. Front Neurosci. 2016;10:196.
24. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65.
25. Joshi CN, Chapman KE, Bear JJ, Wilson SB, Walleigh DJ, Scheuer ML. Semiautomated spike detection software persyst 13 is noninferior to human readers when calculating the spike-wave index in electrical status epilepticus in sleep. J Clin Neurophysiol. 2018;35(5):370–4.
26. Fawcett T. An introduction to ROC analysis. Pattern Recognition Lett. 2006;27:861–74.

27. Benbadis SR, Tatum WO. Overintepretation of EEGs and misdiagnosis of epilepsy. J Clin Neurophysiol. 2003;20:42–4.
28. Benbadis SR. Errors in EEGs and the misdiagnosis of epilepsy: importance, causes, consequences, and proposed remedies. Epilepsy Behav. 2007;11:257–62.
29. Benbadis SR, Lin K. Errors in EEG interpretation and misdiagnosis of epilepsy. Which EEG patterns are overread? Eur Neurol. 2008;59:267–71.
30. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22:209–12.
31. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12:153–7.
32. Marsaglia G, Tsang WW, Wang J. Evaluating Kolmogorov's distribution. J Stat Softw. 2003;8(18):1–4. https://doi.org/10.18637/jss.v008.i18
33. Lodder SS, van Putten MJ. A self-adapting system for the automated detection of inter-ictal epileptiform discharges. PLoS One. 2014;9(1):e85180.
34. Kural MA, Aydemir ST, Levent HC, Ölmez B, Özer IS, Vlachou M, et al. The operational definition of epileptiform discharges significantly improves diagnostic accuracy and inter-rater agreement of trainees in EEG reading. Epileptic Disord. 2021.
35. Aanestad E, Gilhus NE, Brogger J. Interictal epileptiform discharges vary across age groups. Clin Neurophysiol. 2020;131(1):25–33.
36. Chadwick D. Diagnosis of epilepsy. Lancet. 1990;336:291–5.
37. Uldall P, Alving J, Hansen LK, Kibaek M, Buchholt J. The misdiagnosis of epilepsy in children admitted to a tertiary epilepsy centre with paroxysmal events. Arch Dis Child. 2006;91:219–21.
38. Asano E, Pawlak C, Shah A, Shah J, Luat AF, Ahn-Ewing J. The diagnostic value of initial video-EEG monitoring in children – review of 1,000 cases. Epilepsy Res. 2005;66:129–35.
39. McBride AE, Shih TT, Hirsch LJ. Video-EEG monitoring in the elderly: a review of 94 patients. Epilepsia. 2002;43:165–9.
40. Ferrie CD. Preventing misdiagnosis of epilepsy. Arch Dis Child. 2006;91:206–9.
41. Lafrance WC Jr, Benbadis SR. Avoiding the costs of unrecognized psychological nonepileptic seizures. Neurology. 2006;13(66):1620–1.
42. Basiri R, Shariatzadeh A, Wiebe S, Aghakhani Y. Focal epilepsy without interictal spikes on scalp EEG: a common finding of uncertain significance. Epilepsy Res. 2019;150:1–6. https://doi.org/10.1016/j.eplepsyres.2018.12.009
43. Suzuki M, Jin K, Kitazawa Y, Fujikawa M, Kakisaka Y, Sato S, et al. Diagnostic yield of seizure recordings and neuroimaging in patients with focal epilepsy without interictal epileptiform discharges. Epilepsy Behav. 2020;112:107468. https://doi.org/10.1016/j.yebeh.2020.107468

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Kural MA, Jing J, Fürbass F, Perko H, Qerama E, Johnsen B, et al. Accurate identification of EEG recordings with interictal epileptiform discharges using a hybrid approach: Artificial intelligence supervised by human experts. Epilepsia. 2022;63:1064–1073. https://doi.org/10.1111/epi.17206