University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Janmohamed, M;Nhu, D;Kuhlmann, L;Gilligan, A;Tan, CW;Perucca, P;O'Brien, TJ;Kwan, P

Title:
Moving the field forward: detection of epileptiform abnormalities on scalp electroencephalography using deep learning - clinical application perspectives

Date:
2022-01-01

Citation:
Janmohamed, M., Nhu, D., Kuhlmann, L., Gilligan, A., Tan, C. W., Perucca, P., O'Brien, T. J. & Kwan, P. (2022). Moving the field forward: detection of epileptiform abnormalities on scalp electroencephalography using deep learning - clinical application perspectives. Brain Communications, 4 (5), https://doi.org/10.1093/braincomms/fcac218.

Persistent Link:
https://hdl.handle.net/11343/320326

License:
CC BY

BRAIN COMMUNICATIONS

REVIEW ARTICLE

Moving the field forward: detection of epileptiform abnormalities on scalp electroencephalography using deep learning—clinical application perspectives

Mubeen Janmohamed,[1,2,3] Duong Nhu,[4] Levin Kuhlmann,[4] Amanda Gilligan,[5] Chang Wei Tan,[4] Piero Perucca,[1,2,6,7] Terence J. O'Brien[1,2] and Patrick Kwan[1,2]

The application of deep learning approaches for the detection of interictal epileptiform discharges is a nascent field, with most studies published in the past 5 years. Although many recent models have been published demonstrating promising results, deficiencies in descriptions of data sets, unstandardized methods, variation in performance evaluation and lack of demonstrable generalizability have made it difficult for these algorithms to be compared and progress to clinical validity. A few recent publications have provided a detailed breakdown of data sets and relevant performance metrics to exemplify the potential of deep learning in epileptiform discharge detection. This review provides an overview of the field and equips computer and data scientists with a synopsis of EEG data sets, background and epileptiform variation, model evaluation parameters and an awareness of the performance metrics of high impact and interest to the trained clinical and neuroscientist EEG end user. The gold standard and inter-rater disagreements in defining epileptiform abnormalities remain a challenge in the field, and a hierarchical proposal for epileptiform discharge labelling options is recommended. Standardized descriptions of data sets and reporting metrics are a priority. Source code-sharing and accessibility to public EEG data sets will increase the rigour, quality and progress in the field and allow validation and real-world clinical translation.

1 Department of Neuroscience, Central Clinical School, Monash University, Melbourne, VIC 3004, Australia
2 Department of Neurology, Alfred Health, Melbourne, VIC 3004, Australia
3 Department of Neurology, The Royal Melbourne Hospital, Melbourne, VIC 3050, Australia
4 Department of Data Science and AI, Faculty of IT, Monash University, Clayton, VIC 3800, Australia
5 Neurosciences Clinical Institute, Epworth Healthcare Hospital, Melbourne, VIC 3121, Australia
6 Department of Medicine, Austin Health, The University of Melbourne, Melbourne, VIC 3084, Australia
7 Comprehensive Epilepsy Program, Department of Neurology, Austin Health, Melbourne, VIC 3084, Australia

Correspondence to: Mubeen Janmohamed
Central Clinical School, Alfred Centre, Monash University
55 Commercial Road, Melbourne, Australia
E-mail: mubeen.janmohamed@monash.edu

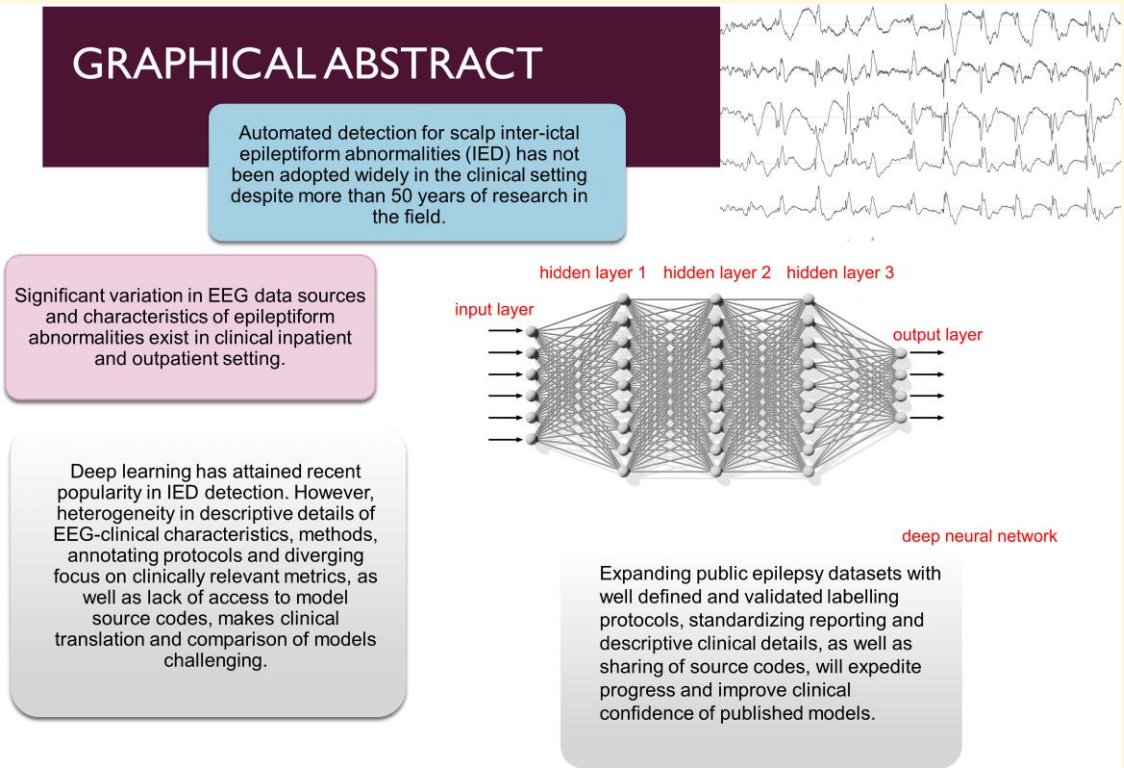
Keywords: EEG; epileptiform abnormalities; automated detection; deep learning; epilepsy

Abbreviations: AUC = area under the curve; DL = deep learning; ICU = intensive care unit; IEDs = interictal epileptiform discharges; IFCN = International Federation of Clinical Neurophysiology; IRA = inter-rater agreement; ML = machine learning; TUH = Temple University Hospital

Received November 07, 2021. Revised May 25, 2022. Accepted August 25, 2022. Advance access publication August 29, 2022
© The Author(s) 2022. Published by Oxford University Press on behalf of the Guarantors of Brain.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

Graphical Abstract



Introduction

Research in computer-assisted automated detection of inter-ictal epileptiform discharges (IEDs) transpired in the decades after EEG acquisition systems became available in clinical practice. The goal was to computerize detection of the 'sharp-transient' hallmark in epilepsy patients.[1,2] An early study pursuing this goal was done in the early 1970s,[3] where a now antiquated computer (PDP-12) was used to discriminate a waveform from a moving average derived from similar polarity amplitudes of 128 preceding waveforms. An indicator pulse was generated when the difference of a waveform amplitude reached a critical ratio. From that time onwards, modern research has explored quantitative time–frequency algorithms as well as machine learning (ML) strategies to develop mathematical models with the intent to achieve reliable automated IED detection.[4] A range of methodologies have been employed to date often in combination, including template matching, autoregressive methods, mimetic analysis, power-spectral analysis (fast Fourier transform, Hilbert and Walsh transform), wavelet analysis, independent component analysis methods and neural network ML methods.[5] However, these methods were tested only on small data sets,[6] thus resulting in low generalizability. Larger data sets improve model performance but require time-consuming feature engineering process.[7] Deep learning (DL), a relatively young field within ML, opens up a possibility to implement modern computing power to detect IEDs and improve workflow efficiency in EEG laboratories. DL differs from traditional ML by using multiple mathematical functions and has the advantage of automating latent feature extraction rather than manual feature selection, making the supervised aspect of training and learning from large data sets more efficient.[7]

A great deal of enthusiasm has been raised regarding DL outperforming expert specialists in healthcare diagnosis and clinical decision-making, and a considerable amount of diagnostic, prognostic and treatment-based ML experimentation has been pursued and published across medical subspecialties. These studies continue to make headlines in various fields. As an example in skin lesion detection, the classification of lesions into melanoma versus benign nevi has shown convolution neural networks (CNNs) outperforming dermatologists in dermoscopic examinations.[8] In another landmark study for identifying and grading diabetic retinopathy using retinal fundus photographs, a DL neural network showed above expert-level sensitivity and specificity of over 90% in detecting referable diabetic retinopathy and macular oedema.[9] This required labelled imaging by 54 ophthalmologists on a data set of 128 000 images for training and validating and testing in a subsequent data set where it outperformed health experts.

Such examples of remarkable success however should not be prematurely taken to conclude that ML in health has

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

reached an implementational level in real-world clinical practice. When the above promising retinal classification model was deployed in a real-world prospective study in Thailand, several impediments were identified affecting system performance.[10] Twenty-one per cent of the retinal photographs were rejected by the algorithm as they did not meet the system's high standard for grading even when they were of adequate quality to be graded by the human visual eye. Real-world clinical data on the ground are frequently affected by a diverse range of technicalities which health experts have to regularly deal with, and this is particularly pertinent in the EEG and epilepsy world.

DL and ML in the EEG field covers a broad scope of research including epilepsy, sleep diagnostics and brain–computer interfacing.[11,12] Within clinical epilepsy itself, ML approaches have been investigated for seizure detection,[13,14] seizure prediction,[15] epileptiform detection,[16] epilepsy imaging, genetic mining and classification, medical[17] and surgical treatment decision-making and clinical outcome prediction.[18] High discriminative abilities have been asserted in these varied fields; however, there remains an uncertain perspective of real-world implementation and generalizability.

A recent study related to EEG IED detection employed a 10-fold cross-validation method on over 13 262 IED candidate waveforms.[19] A very impressive area under the curve (AUC) of 0.98 of IED detection was cited for a DL model developed and termed as SpikeNet. Additionally, an AUC of 0.847 was also reported for classifying whole EEGs using a binary classifier trained using 10 extracted features. The model reportedly outperformed fellowship-trained EEG experts to detect individual IEDs. This number, however, needs to be contextualized. All data were obtained from a single centre including training and test data set and an external out-of-hospital test data set was not employed. An epoch-based graphical user interface point and click format (NeuroBrowser) was employed which required blinded classification by the raters. The overall inter-rater reliability in this particular study for these blinded reviewers agreeing on candidates as spikes was only fair with (Gwet κ) of 48.7. Most importantly, the source code for this model has not been available on public repositories to validate on external independent data sets. This external validation limitation in ML is well known.[18]

This review summarizes some of the perspectives of clinicians who have provided clinical support in DL IED detection in collaboration with data scientists via EEG data obtained from tertiary epilepsy centres in Melbourne. The article will allow data scientists and researchers entering the automated IED detection field to quickly understand the basic nature of the EEG data used in epilepsy management, challenges they will encounter upon embarking their journey and recommendations on moving the field forward.

Search strategy and selection criteria

References for this review were identified through searches of PubMed with the search terms 'inter-ictal, 'epileptiform', 'spike', 'deep learning', 'automated software' and 'epilepsy' from 2010 to January 2022. Only papers published in English were reviewed. The final reference list was generated on the basis of originality and relevance to the broad scope of this review.

Why research in this field and limitations

The digital era has opened itself to automating tasks requiring human efforts, especially those which are repetitive and time-consuming (Table 1). This pursuit has been embarked to make hospital workflows more efficient. A routine EEG recording of 30 min usually takes anywhere between 5 min and an hour (median: 13 min) to be visually assessed and reported by an epilepsy specialist, depending on various factors, including presence of abnormalities, length of the EEG and artefacts present.[20] This time can also be increased or decreased based on the setting of the EEG. In the intensive care unit (ICU) setting, 24 h of abnormal continuous EEG being reviewed for only seizure identification required a median of 44 (±20) min in a retrospective review of conventional review versus quantitative EEG comparator study.[21] In contrast, a DL algorithm can take minutes to label and provide prediction labels for a 24 h EEG. A recent paper showed an average computational time of 7 s to label signal lengths of 1 h.[22]

During manual review, Identification and interpretation can take longer for more difficult and complex EEG data. An example would be intracranial data of a patient with a complex epileptogenic zone and several dozens to hundreds of electrode contacts resulting in a vast number of channels to review. Inconsistent labelling is also common in practice as different EEG technicians and clinicians use different approaches and terminology in marking data. A successful computer-assisted detection would theoretically vastly reduce the time and improve quality of labelling done manually by EEG scientists, technicians and clinicians.

Table 1 Pros and cons of future computer-assisted detection in EEG laboratories

Pros
• Speed labelling and substantial data reduction leading to faster workflows
• Substituting unavailable expertise in low-resource countries
• Artificial intelligence is purported to have the potential of better results than traditionally trained experts.
Cons
• Missed true epileptiform discharges (false negatives) with the potential to delay treatment
• Exaggerated labelling of artefacts as abnormalities (false positives) (see Fig. 3)
• Reduction of job and learning opportunities for EEG scientists and epilepsy trainees

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

The most concerning limitation of implementing automated detection in future workflows would be misreporting of an EEG by an overseeing clinician, in particular a non-expert epileptologist, biased by the automated programme. Every EEG has variation, and no model will ever result in a 100% accuracy. An EEG may be reported as positive when not, resulting in unnecessary and even harmful treatments being implemented, and conversely false negatives may delay treatment with the potential to cause harm to patients. This problem has also been often noted in global clinical practice, outside the expert neurophysiology community and addressed in a series of articles in 2013 appearing in Neurology. In a survey of 47 trained neurophysiologists, during the annual meeting of the ACNS in 2010, many noted coming across misread EEGs and 38% encountering them frequently.[23] In a more recent study in India,[24] 1862 EEGs were prospectively performed to identify the prevalence of benign epileptiform-like variants (BEVs). Under recognition and misreporting were common in the neurology community. Amongst 101 subjects whose previous raw EEGs were accessible, 30% of benign variants were noted to be misinterpreted as epileptiform abnormalities. Several recommendations and guidelines across the epileptology literature[23,25,26] have been made to reduce this risk and efforts are in place to increase the training, teaching and reporting of EEGs. In the hands of inexperienced EEG readers, an automated detection programme may confound and potentially worsen misreporting.

Overview of scalp EEG data sets available for automated detection

A wide array of EEG recording types can be retrieved from hospital-based EEG servers (see Fig. 1). A scalp outpatient routine EEG is the simplest of the raw EEG data sets available and is usually recorded in a 10–20 electrode configuration, with or without ear electrodes. Routine EEG recordings are frequently done in rested patients who are not in an unwell clinical state and can generally, at most times, follow instructions. The quality of the EEG signals would be amenable for machine and DL as recording technicians in real-time are able to improve the quality of the signal recording and annotate important segments for a further clinician's review. This has been a common data set used in DL literature. Routine outpatient EEGs typically range between 20 and 30 min and sometimes a more prolonged 1–3 h sleep-deprived or non-sleep-deprived EEG may be requested by the clinician overseeing the patient's care. Sleep-deprived EEG similarly provide good quality recordings given artefacts from movement and muscle are considerably reduced during sleep, and a marked surge in epileptiform abnormalities is seen in both focal and generalized epilepsy during sleep.[27–29] Sleep, however, presents a different overall background from which the epileptiform abnormality emerges, and the epileptiform abnormality can present different morphologic characteristics and of briefer duration in the case of genetic generalized epilepsy.[30]

A downside of routine EEGs is that they are less likely to have abnormalities to gather for the training data set given their shorter duration. However, major tertiary referral centres however would still have several hundred to thousands of routine outpatient EEGs that are abnormal and contain epileptiform abnormalities stored on their servers depending on the protocol of archiving used and format compatibility with modern software. In a hospital, an EEG recording can also occur in a ward-based inpatient setting, a multiple-day elective video monitoring setting or the critical care setting. Video-EEG recordings of patients who are electively admitted for a multi-day recording would be intermediate in quality. Scientists are able to correct loose electrodes and aim to reduce artefact contamination, improve impedances and educate the inpatients to aim for better technical recordings. Sleep background is also available, and video is always available to correlate abnormalities for review. Here the number of electrodes can differ depending on the purpose of the elective admission. Recordings for surgical localization regularly have additional sub-temporal electrodes. Sometimes symmetric or asymmetric higher density electrode coverage in addition to the standard 10–20 electrode placement system may be carried out in different regions of the brain which introduce variation. Such video-EEG recordings can easily be processed to a 10–20 format for further data processing.

Lesser quality data sets would include ambulatory EEG[31,32] recorded when patients are up and about at home, introducing large movement and myogenic artefacts and where an overseeing scientist is not reviewing the record until the leads are removed the next day or at the end of recording duration. Perhaps, most challenging of all would be critical care patients where electrical interference from surrounding equipment causes significant artefacts, electrodes may be placed in non-traditional positions or excluded due to craniotomies and the background may be confounded by sedative medications or the underlying brain insult. Prolonged ICU continuous EEG is often performed in hospitals for monitoring seizure activity of critically unwell patients.[33]

The nature of EEG background and epileptiform discharges

The EEG presents a wide diversity and dynamic nature of both background and epileptiform discharges in an EEG. This variation has to be understood before embarking on the ambition of a universal IED detection model. Figure 2 demonstrates sample EEG epochs of epileptiform variation in a genetic generalized epilepsy data set showing a sample of diversity in epilepsy IEDs for one epilepsy type.

The normal background of an EEG is dynamic and can be divided into normal awake, drowsy and sleep stages.[34] Background frequencies are slower and less dynamic in encephalopathic patients or those with developmental delay and neurodegenerative conditions. There can even be an association of background frequencies with age. Paediatric EEG has a much more complex range of normal background

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

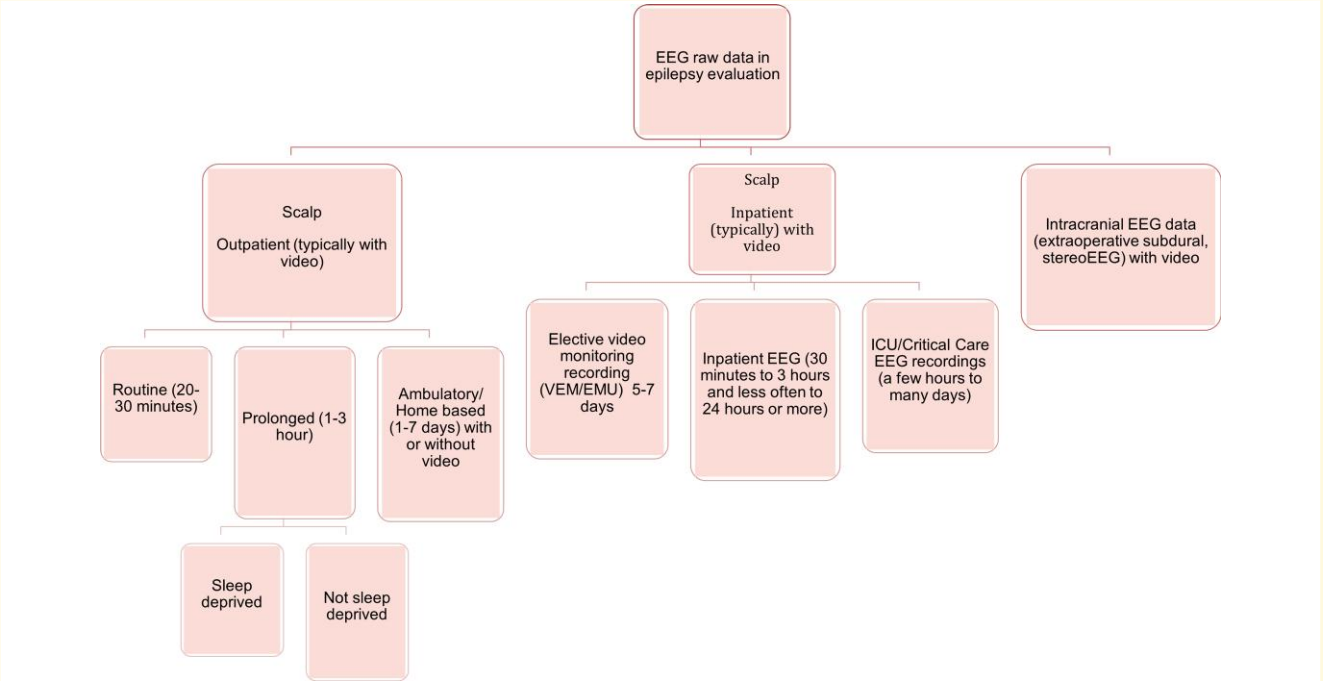


Figure 1 The structure of data available from hospital-based EEG servers.

while older patients demonstrate slowing of the dominant posterior alpha activity. Simple state-occurrences such as eye closure can also modulate the background. A wide multitude of technical issues and artefacts can add to tremendous variation in the background during the awake state and some of these can resemble epileptiform abnormalities.[35] These can occur in both the awake and drowsy state. Several background variants of the normal EEG can occur including alpha variants (fast and slow) and posterior slow waves of youth. Further, BEVs can occur in the EEG during awake and drowsy state and may include Benign small sharp spikes (or Benign Epileptiform Transients of sleep, BETs), wicket waves, 14 and 6 positive spikes, and 6 Hz phantom spike and wave.[34,36,37] Physiologic changes in the drowsy and sleep record include attenuated posterior rhythm, central theta, V-sharp (vertex) waves, large K-complexes, spindles, arousal patterns, positive occipital sharp transients, as well as temporal or diffuse rhythmic theta and high-amplitude delta slowing.[38] Detection models may confuse some of these morphologies with epileptiform discharges (see Fig. 3), especially those discharges showing rhythmic sharply contoured waves forms or even delta duration slow waves, such as the kind seen in spike/slow wave.[39] Similar to dynamic changes in background which characterize a normal or abnormal EEG, there is no uniformity in epileptiform discharges within and across patients. Epileptiform discharges can vary in duration, morphology, periodicity, topography (Fig. 2) and can be modulated by either state changes or other provocative manoeuvres. The diversity of epileptiform discharges can include spikes (20–80 ms), sharp waves (80–200 ms), spike–slow wave, sharp-slow wave, polyspikes, polyspike–slow waves, as well as fast activity associated with the aforementioned. These can be located, within one subject, in one region or in two or more regions either in one hemisphere or bilaterally. When these engage bilateral networks, they are referred to as generalized epileptiform abnormalities and when confined to one hemisphere in a few electrode sensors as focal abnormalities. They can occur as isolated transients or be part of a sequential train or run which can be periodic[40] or quasi-periodic. In some patients, these frequently recur through the duration of an EEG recording but may only occur only occasionally in briefer recordings. They can occur in combination with any of the background states mentioned above.

Sleep modulates and accentuates the occurrence of epileptiform discharges[30] in focal, as well as in generalized epilepsy and so does provocative manoeuvres which in different epilepsies may include photic stimulation,[41] hyperventilation or even things such as visual, audio or cognitive tasks.[42,43] An ambitious all-spike (universal IED detecting) DL model should have undergone training with a vast amount of EEG capturing the majority of this variation. The differences further are confounded by EEGs available from different settings, where different noise levels will be present. To reiterate, an outpatient ambulatory EEG recording is not equivalent to a routine resting EEG in terms of background quality and both of those will be different to EEGs acquired in a critical care EEG.

Inter-rater agreements and the gold standard comparator

One of the main concerns in current DL and ML studies is the lack of an unambiguous framework of what the gold

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

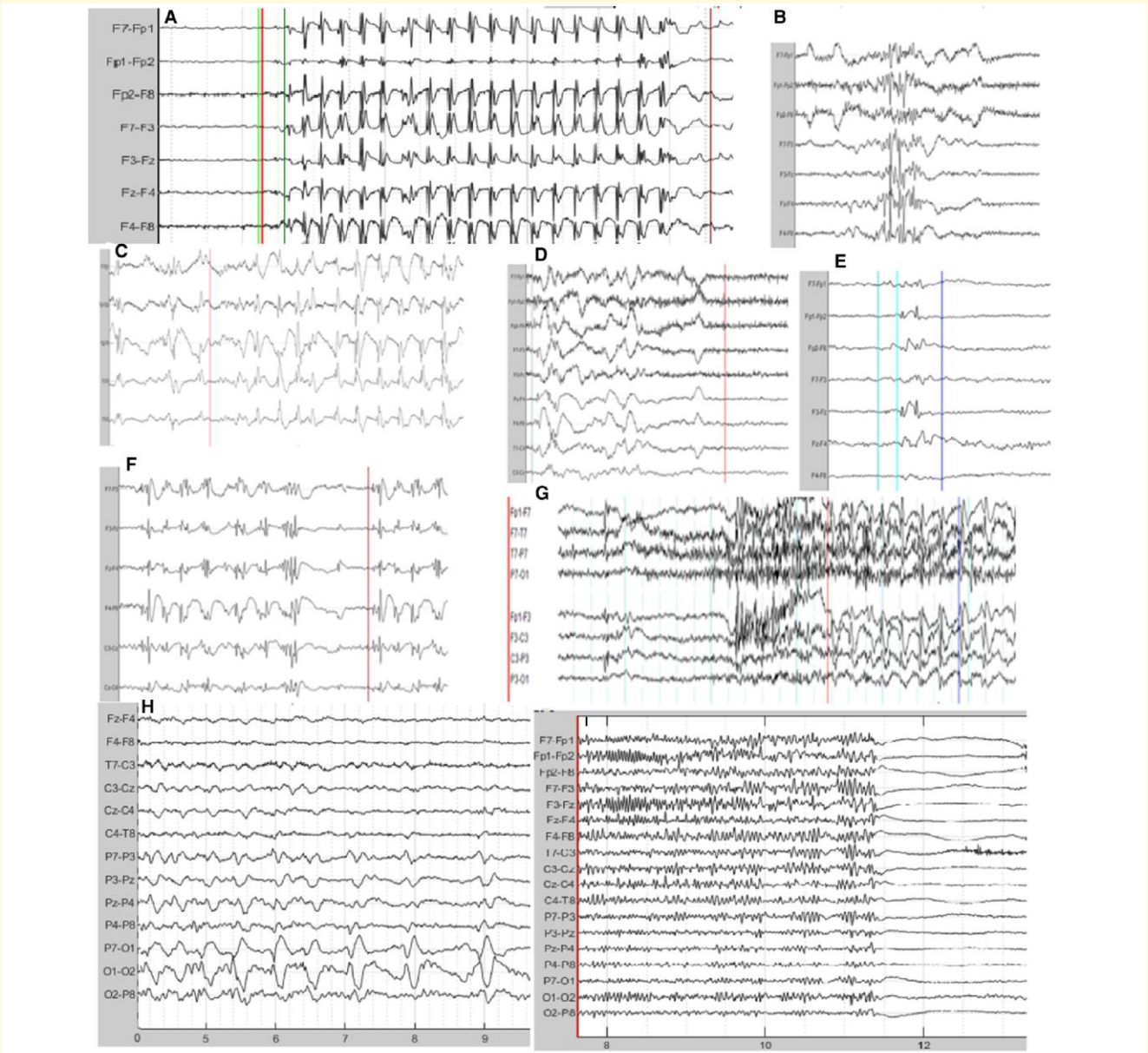


Figure 2 Epileptiform variation in Genetic Generalized Epilepsy EEG data sets. (From left to right) (**A**) Classic 3 Hz spike and wave on transverse montage, (**B**) polyspikes with EMG artefact in frontopolar channels and eye movements, (**C**) slow spike/wave on transverse montage, (**D**) mild EMG affecting frontal channels with embedded small spike and waves and irregular slow waves, (**E**) fragments on transverse montage, (**F**) polyspike/slow waves on transverse montage, (**G**) marked EMG artefact confounding epileptiform abnormality in temporal and frontal channels on longitudinal montage, (**H**) a train of focal posterior sharp waves and a (**I**) generalized paroxysmal fast burst.

standard or 'ground-truth' is for determining the accuracy of the final computing model.[44] Given the consistent real-world underperformance in the reliability of computers versus humans in EEG IED discretion, the reference standard today remains visual review and classification by a trained epileptologist or clinical neurophysiologist. This presents some limitations as inter-rater reliability amongst EEG readers has been controversial due to the perceptual phenomenon and probabilistic art of reading spikes.[45] For decades, the question of inter-rater agreement (IRA) has been investigated in parallel with the question of computerized detection. The Food and Drug Federal Administration requires three electroencephalographers (EEGers) in the annotation process for approval of an algorithm.[46] Some authors have directly looked at inter-rater reliability and specify the number of EEG raters selected for an internal criteria of a definite IED.[19] Halford[5] mentions at least four different agreement criteria used by ML authors for referencing actual IEDs in his review paper including concordance between all raters, a cut-off number of raters from the whole group, reconciled

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

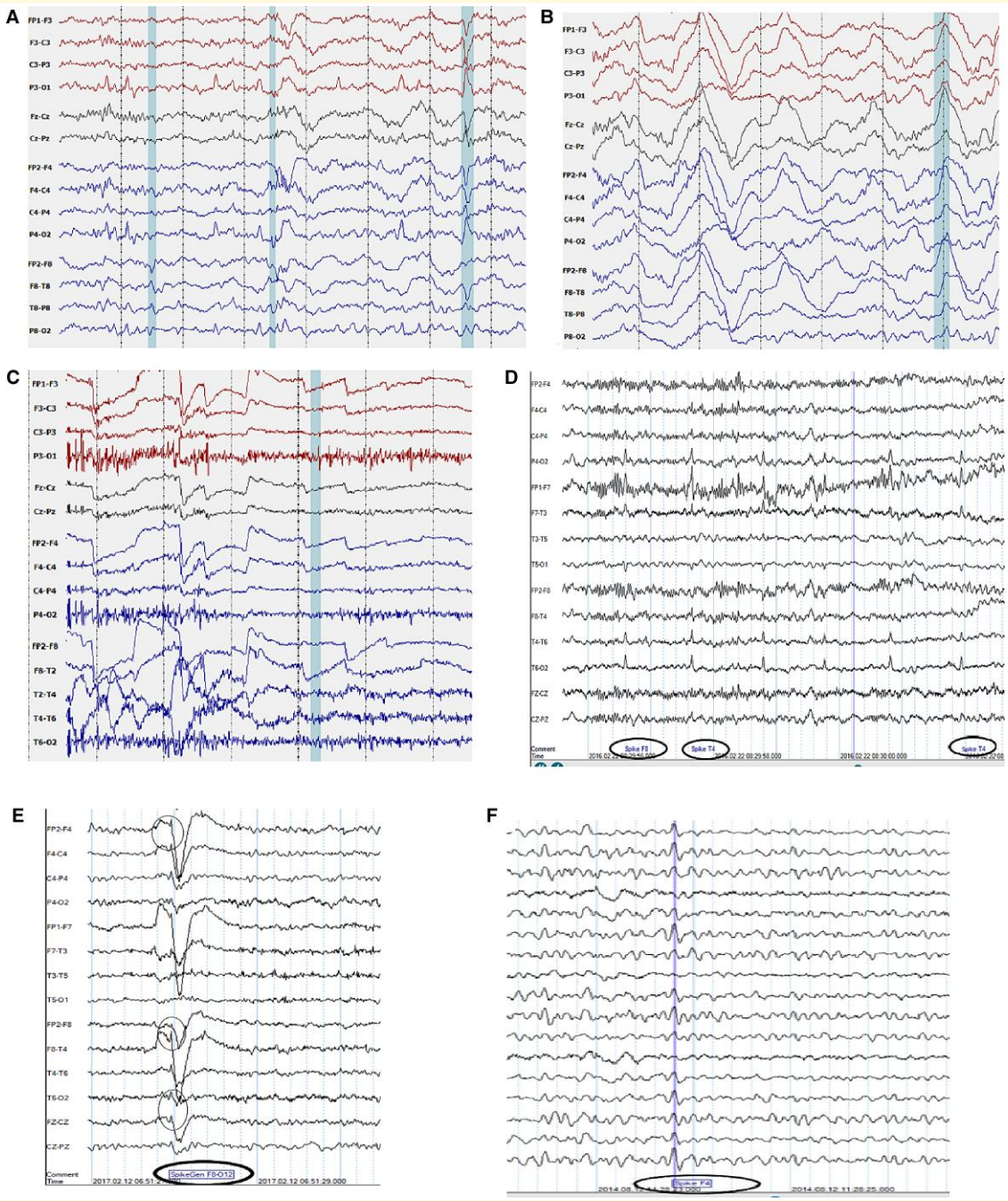


Figure 3 Artefacts mimicking interictal epileptiform abnormalities. IED mimics (**A**) *v*-wave mimicking sharp wave and is labelled as abnormal by algorithm. (**B**) High-amplitude slow wave in Stage 3 sleep causing false positive, (**C**) ocular artefact, (**D**) ECG artefact picked up as runs of IEDs, (**E**) lateral rectus spikes and (**F**) wicket spike picked as false positive.

rating amongst reviewers and some papers using only one rater to define the gold standard. Expert pooling has been found to be better and larger group sizes from 3 to 10 have been reported to be ideal yet judiciously selecting expert EEGers for IED annotation research projects may reduce the need for this number.[46]

Although the above may lead to a perception that IRA amongst EEGers is imperfect and unreliable, this is not entirely accurate. Several studies have shown moderate to substantial IRAs with some studies report higher Gwet or kappa as well as high performance of blinded clinical experts compared to an unblinded gold standard.[47–49] IRA in 'whole EEG' categorization remains high given low-perception spikes are contextualized by EEGers before concluding the report as normal and abnormal. A limitation of poor or fair IED agreement studies is that reviewers are blinded to clinical context or

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

deal with very short segments of data when tested.[50] Most real-word EEG review of waveforms requires an awareness of patients' age, compliance with recording instructions, technical quality, sedative agents, pharmacologic agents, previous EEG characteristic, clinical context and the conscious state of the patient during recording.

A recent paper provides a good benchmark and sheds light on what can be used as a gold standard. Kural *et al.*[47] assessed six criteria that are used to determine what is or is not an epileptiform abnormality to assess inter-rater variability amongst clinicians for each feature criteria and assess the International Federation of Clinical Neurophysiology (IFCN) criteria as a whole for validity. In the study, they used a strict methodology to confirm an IED. It required two reviewers agreeing that the candidate waveform was a sharp-transient and furthermore additional criteria of a patient not only to have confirmed epilepsy but of the selected transient being concordant with the patient's recorded seizure and location as expected in that syndrome or focality (interictal, ictal and syndrome correlation). The clinical context was thus extensively incorporated in the decision-making of what is or is not a spike thus setting an acceptable benchmark for gold standard for the purposes of that study. With that gold standard other methods of defining epileptiform were evaluated. Blinded reviewers utilizing four or five of the six IFCN criteria together provided a strong accuracy of the waveform being labelled as epileptiform with accuracy levels of 91% (95% CI: 83.6–95.80) and 88% (95% CI: 80–93.6) against the gold standard. Furthermore, experts solely using their clinical experience with no protocol method and simply consensus provided a 92% (95% CI: 84.8–96.5) accuracy. It is important to note that all these experts were blinded to the original two reviewer unblinded gold standard assessment. The six IFCN criteria used included the morphology of spikiness or sharpness, asymmetry of ascent–descent slope, duration difference from background, an after-going slow wave, background disruption and a concordant voltage map.

Assessment of current DL studies

A systematic review by the authors submitted and under review examined 17 recent DL scalp EEG studies for IED detection (Supplementary Table 1). Of the studies, 60% used focal and generalized epilepsy EEGs, whereas the remaining focused on either or a BECT data set. Six studies used data from more than one centre for testing set. Two papers used more than a thousand EEG recording data set with median number of EEG recordings per study as $n = 166$. Routine EEG recordings were most commonly employed whereas prolonged recordings next most common. All studies utilized supervised or semi-supervised learning requiring labelled data; however, significant variation in gold standard identification of IEDs existed. Montage applications were mentioned in some studies and not mentioned in others with the most common channel combination for data input as longitudinal bipolar montage. The most common architecture employed for DL was CNN and long short-term memory with combinations and hybrid comprising the rest. Accuracy-based measures AUC/balanced accuracy or simply accuracy were reported in all studies; however, sensitivity in combination with either false positives or precision was not reported in some studies. Overall, a wide variety of data preparation, pre-processing methods and neural architecture techniques were utilized. Presentation focus of performance metrics vary significantly amongst studies introducing difficulty in comparisons of strength of models. Further technical details of the data properties, pre-processing methods, design of DL architectures and layers, optimizers and pooling methods can be reviewed from the original papers and a systematic review by the authors is under review.

Annotation standards

Recent literature on DL IED detection do not usually elaborate on the clinical annotation protocol used for the supervised learning process and we found this information lacking in publications. A six-way labelling classification of epilepsy EEGs is employed by the Temple University group which has availed their public data set on the internet.[51] Proper annotation labelling may be important for the performance of the algorithm. An abnormality can be marked using a single marker or a start and end marker. The windowing method employed for training and testing purposes for IED detection may incorporate partial normal segments in the windows designated as abnormal containing epileptiform abnormalities. If only one marker is used it can be placed at various points along the discharge most frequently at the negative peak of a selected spike portion. No strict rigour can be employed here due to the vast heterogeneity of how transients and prolonged discharges appear. Even when a more laborious 'start' and 'end' markers are utilized there is frequent inter and intra-rater variation from our centre's experience in labelling as abnormalities often do not have clean onsets and offsets. A decision may be made to annotate the first spike onset, however spikes may terminate before the discharge has ended in the case of generalized epileptiform abnormalities.[52] Conversely, discharges may emerge with some abnormality in background or rhythmic slowing before showing clear spike morphology. In the case of stereotyped repeated focal transients, the onset and offset may be easier to define. An annotation marker may be generic with an instant timestamp without regard to the channels involved, or it may be specific and labelled according to the specific channels involved. The Temple University Hospital (TUH) events public corpus has made an effort to label abnormalities based on specific channels involved and may allow more precise abnormal signal input for the subsequent learning process.[53] These annotations of the data sets however will need to be systematically validated.

Metrics evaluation

Utilizing sensitivity, specificity and AUC of each model does not always translate into useful clinical assessment. There

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

remains a challenge with reporting metrics in regard to IED DL models, and the most clinically useful metrics are variedly reported in the currently existing literature. Both accuracy and to some extent AUC as understood by data scientists in the field, unless contextualized with other metrics, can be misinterpreted by clinicians. This has been noted by some authors.[54] Summaries of recent DL models consistently report over >0.9 or 90% AUC[19,55–59] A model can be stated as having 90% 'AUC' or accuracy and yet be unreliable from the clinical perspective. This will occur if the normal windows for the majority part were correctly predicted even if the abnormal discharges, which occupy very brief lengths and occur sparsely, were all mostly missed. The only tuning of the model to give a high accuracy would be to reduce the number of false positives which could be attained by raising the detection or perception threshold (low sensitivity setting). True negatives in such a model will be high in both the numerator and denominator, falsely giving an 'accurate performance'. An AUC does not always solve this problem as both a high sensitivity and specificity do not necessarily address the issue of false positives (examples in Fig. 3). This problem of imbalanced data set during ML and DL training occurs when normal background input vastly exceeds abnormal windows with a ratio of up to 1:1000.[56] Data augmentation methods have at times been incorporated to increase the representation of the spike minority class using oversampling techniques.[60,61]

When evaluating a DL model, several other metrics, therefore, have to be taken into account and few metrics are helpful as standalone measures to give a perspective on success (see Table 2). Precision reports true positive IEDs in the entire set of predicted IEDs. It represents the positive predictive value in clinical terms. A higher precision implies a lower false positive rate. False positives per minute or hour is a simple and informative metric which provides accurate insight into performance when full-length EEGs are evaluated in the test data set. Precision or false positive rate coupled with sensitivity represent a better measure than specificity and AUC. Class imbalance as described above skews specificity TN/(TN + FP) and other measures dependent on it like AUC. Amongst other parameters of high utility is the F1-score which provides a weighted average of both sensitivity and precision and the AUPRC which is the area under the precision–recall curve (AUPRC). The F1-score weights the two most important variables and will take into account false positives and false negatives without contaminating or exaggerating performance with the imbalanced true negatives.

Table 2 Performance metrics commonly used in deep and machine learning studies

Metrics of clinical utility for IED detection
• Sensitivity: Proportion of true gold standard IEDs correctly detected
• Precision: The proportion of true marked gold standard IEDs to all machine predicted positive labels. (True positives)/(true positives + false positives)
• False positive rate: Rate of false positives which were not classified by the gold standard as IEDs typically reported in per hour
• F1-score—This takes into account the two most relevant metrics of precision and recall.
• AUPRC—Area under the precision–recall curve (AUPRC) which differs from the area under the ROC curve. A model achieves perfect score when it identifies all epileptiform abnormalities without marking normal or benign abnormalities
Metrics of limited clinical utility in isolation
• True negatives, specificity, accuracy and AUROC (area under ROC curve)

Channel-based labelling has its advantage. For focal abnormalities and artefacts maximal involvement is in few channels can improving the target specificity for data training. On the other hand, if one maximally involved channel is selected for training this has the risk of ignoring data from the remaining field extent of the IEDs. If per channel evaluation is desired modifications to appraisal will be required in a mixed unselected data set as generalized epileptiform abnormalities or spikes with broad fields will show preferential biasing compared with localized spikes. From the clinicians' perspective, single-channel annotation would be tedious to review, present redundant data and pose difficulty in evaluation when many channels are involved. A single timestamp even for generalized or broad field focal abnormalities would serve the intended purpose of automated detection as a screening tool.

We found a lack of clarity in many papers as to how redundant predictions for a single contiguous discharge are dealt with. Continuous EEG data are frequently segmented into short windows (ranging from 0.5 to 2 s) and is trained or tested using a sliding window of overlapping or non-overlapping windows (50–75%). From a clinician's perspective, redundant (repeated) markings between pre-labelled start and end marking of a contiguous epileptiform abnormality should not be counted as true or false positives and should not influence metric calculations. This may differ from the understanding of the detection target in which IEDs are thought exclusively as 'spikes'. We reiterate in this article the many different types of epileptiform patterns including bi- or triphasic sharps, spikes, polyspikes, fast patterns with or without slow waves occurring in isolation or prolonged repetitive bursts exceeding conventional window segmentation (see Fig. 2) used in DL are common.

In the case of highly active prolonged abnormal EEGs, the recording can be trimmed to reduce margin of error for the manual annotator instead of using dozens of hours. Furthermore, manual review and validation of the automated spike labels should also be performed to ensure any extras detected by the ML are false positives and not detected but unlabelled true positive IEDs.

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

Whole EEG classification

In this area of interest, some authors have reported on IED-free versus not-free to categorize whole EEG classification.[19,57,59] This may implement a IED rate threshold to categorize EEGs into normal versus abnormal which differs from criteria used in real-world classification by epileptologist. Clinically, one unequivocal epileptiform abnormality suffices to classify an EEG as abnormal. If EEGs were sorted from clinical reports, caution needs to be adopted given other EEG features, most notably focal slowing or indeterminate findings, can lead to the classification of an EEG as abnormal without an epileptiform abnormality. Textual mining of reports for data set retrieval should specifically require the presence of epileptiform abnormalities and the conclusion of the report should be ascertained. Sensitivity and specificity calculations are easier to calculate for whole EEG classification as actual positives and actual negatives are easier to define without the challenge of defining windows in relation to IED durations. IRA on whole EEG classification is higher than for individual IEDs.

Future directions and overcoming the IED challenge

In seizure detection, low false alarm rates (<1/h) and high detection rates (70–80%) have been achieved,[62] and software are operational in some hospitals for seizure alerting, detection and continuous monitoring.[63] On the other hand, despite more than 50 years of study in this study area, commercial or open-source software have not become pervasive in clinical use for the detection of IEDs on EEG recordings despite the immense benefit to time and labour challenges in an EEG laboratory. A recent commercialized spike detection software trained using DL algorithm, Encevis Solutions[64] (Austria), sensibly uses clustering to overcome the high rate of false detections 112/h for a high sensitivity of 89%.[54] Persyst 13 has been reported in one study as non-inferior in performance to senior EEG technologists[65] at a low-perception predictive setting (high sensitivity setting) but was found to have much higher false positive rates at various perception thresholds compared with board-certified EEGers reviewing Epoch-based transients.[66] It remains to be seen how well the software will perform on more varied, larger and longer unselected EEG data sets using a reliable gold standard. There remains significant scepticism amongst EEG technicians and clinicians as to the benefits of available software accurately guiding the process of capturing and labelling spikes on scalp EEGs and subsequently quantifying IEDs let alone precisely making a judgement on classification of a scalp EEG into normal or abnormal (IED-free versus IED-EEGs).

The scepticism and difficulty in readily available DL algorithms for computer-assisted clinical EEG reporting has been due to a great number of barriers. The wide heterogeneity of methods, statistical approaches and reporting in current literature introduces difficulty in comparison of models. The models may appear accurate in a single-centre data set but their applicability to multiple data sets is more challenging. Standardized descriptions of data sets and reporting metrics is essential remains a priority in this field.

Proposed hierarchy for gold standard epileptiform detection

Given no consensus exists on the gold standard to be used for identifying and labelling epileptiform abnormalities for training and testing DL models, different approaches can be utilized based on extent of clinical support available in the centre. In the most reliable situation, epilepsy experts (even a few) assess each spike or sharp wave in relevant time epochs as an epileptiform abnormality having awareness of the context of the patient's profile (clinical history and imaging) and have access to longer EEG recording or at least more than just short epochs, including ictal patterns and locations to correlate for concordance. EEG reviewing and reporting, especially for inpatient EEGs, eventually considers all clinical comments regarding why the EEG is performed. This criterion may not be practical for big data research projects. Second, despite the above-mentioned limitation, few sufficiently trained EEGers utilizing the four or five of the six IFCN criteria mentioned above, preferably with basic context for the EEG being labelled, would be a validated method.[47] The third method would be epilepsy experts either marking via 'experience' without adequate exposure of the entire EEG and clinical context of the respective EEGs. Due to sub-par inter-rater concordance frequently cited in the literature, the third method remains controversial in terms of how many people should agree. Occasional papers, however, continue to show an adequate level of expert agreement, even when this method is employed.[47,48,67]

If there are limitations in obtaining epilepsy experts to annotate and determine ground truths, a layered approach can be implemented where a less rigorous method is employed to annotate the training data set and a more rigorous method used to annotate the testing data set. This will ensure that the performance results provided in the study have been compared against an adequate gold standard in that centre.

Standardizing reporting of methods and results

A recent exemplary paper provides details to properly understand a DL publication of IED detection. Adequate description and division of EEG data was provided, epilepsy syndrome details, method and algorithmic details and most importantly comprehensive performance metrics results.[22] For the data set, there should be a clear explanation of the recordings being either scalp or intracranial, the environmental setting in which the EEG was performed, and the breakdown of the type of EEGs used for both training and testing data sets. Additional clinical characteristics of cohort can be helpful. Extended labels are preferred to single

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

timestamps during manual labelling otherwise the IED discharge will be assumed to be the window size by the performance evaluator to enable calculation of relevant metrics. If combinations of heterogenous recorded EEG data in different inpatient and outpatient settings are used, the proportion of different respective EEG types implemented in both the training and testing data set should also be described. Electrode configurations used and channel derivation from electrode montaging should be mentioned as these can introduce some differences in signal characteristics. Average referenced signals can be different depending on vertex or ear electrode referencing or all-electrode averaging which can be different from bipolar or transverse derived signals. The montage ultimately chosen for testing and training may introduce some variation in the derivative signal and may not translate well into a test data set using a different montage configuration for signal derivation. As an example, waveforms may appear sharp and phase reversing on bipolar whilst not appearing different from the background in an average montage. Similarly averaging can sometimes bring out waveforms which undergo differential voltage cancellation in bipolar montage due to equal strength amplitudes. In a recent study, we found combined training on transverse and longitudinal montages simultaneously provided high F1-score in a recent GCN convolution model.[39]

Very few publications provide details on epilepsy type, syndromes and nature of discharges and whether the bulk of the abnormal discharges or the majority background used was derived from awake or sleep state or in what estimated proportion. This can be important as a data set used to train a focal epilepsy model may not be appropriate for a generalized epilepsy test data set. Furthermore, some models may work well on awake background but present false positives in sleep EEG due to low frequency waveforms being confused with slow wave abnormalities.

Evaluation results are frequently present on test data sets in a summarized pooled manner. A few outlier EEGs causing poor performance may markedly skew the results to show the algorithm as inaccurate whereas this may be the case because of only a few EEGs in which the algorithm failed significantly. In our ongoing work evaluating an unpublished data set[39] implementing a Graph convolution method (viewing an EEG montage as a graph theory using electrodes as nodes and pair linkage as edges), we found removal of 4 outlier EEGs from a test data set of 28 EEGs markedly improved precision at a 0.80 detection threshold from 28 to 63% with only a 10% drop in sensitivity (errant spike windows predicted reduced from 781 to 57). Outlier identification efforts, although time-consuming, should therefore be made in conjunction with a trained epileptologist to find the reasons why the overall results of an algorithm may be poor.

Standardized metrics should be reported including as many performance metrics as possible to provide a holistic view rather than a focus on accuracy or AUC. Most importantly and invariably clinical useful metrics of false positive rates and sensitivity should be reported within abstracts and conclusions. Other details on how the analysis and performance statistics were calculated should be elaborated in forthcoming DL studies. Were brief epochs or entire lengths of EEGs evaluated in the test data set to decide on prediction accuracy? Accuracy/AUC calculated on IEDs in a substantially imbalanced real-world data set of whole EEG recordings is different from Accuracy/AUC calculated on a data set limited to segmental review or epochs with an attempt to balance normal and abnormal segments in the test set. Selection bias can also be introduced into the epoch-based methods as noise-free segments with better technical quality, uncontroversial epileptiform discharges and more normative backgrounds with less complexity can be chosen by the data set retrieving team. The problem of imbalanced data set has to be tackled in this field as epilepsy data will always have the vast majority >95% or more of its signal to be normal background apart from a few outlier intractable epilepsy patients who have frequent, near continuous or continuous epileptiform abnormalities interspersing background. The benchmarking test data set must therefore be an imbalanced data set if any real-world clinical utility is to be desired.

Public data sets and source code-sharing

Cross-testing is vital and will reveal the actual performance of a model. This has only started to be employed in seizure detection. False detections of seizures ranged from 0.15 per hour to 2.5 per hour depending on different data sets used.[68] The hospital and locality ethics of sharing EEG data makes it complex for potential collaborators seeking to implement their model algorithms on external data sets. Multi-centre data set collation should nevertheless continue to be pursued. Epilepsy centres collaborating will be able to reach target numbers reached in imaging classification by share loading contributions and be able to allow the DL model to be trained on a large amount of morphologic, topographic and artefactual variation of windows containing epileptiform discharges. This cannot be done without a collaborative mindset.

Source codes detailing current DL models being experimented and published in the automated IED literature are not available for other researchers to replicate on their own data set and subsequently critique, improve or even compare with their own planned models. This very likely may be due to researchers considering that their models could be improved to a point of commercialization or alternatively suggest a lack of confidence on the generalizability of the model and thus keeping model details restricted and neural networks architecture explained in a general way. Such source code-sharing has been done in seizure detection algorithms.[69] Any researcher who has published a model should avail their source code on a public repository to allow people to quickly test and validate the stated model performance on their respective private data sets. This will allow robust peer review. Such feedback can be provided back to the publishing author who can further fine-tune his model or be

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

made aware of the model's performance on different data sets to his. This would be easier than researchers sharing or requesting EEG data from other centres. If source code-sharing for cross-testing is not desired the next step would be standardized data sets to be made available in the public domain against which models from different research groups can be tested and compared. This however will not allow peer review of performance as model testing is carried out by the same authors and selection bias and selective reporting can still result. Thus, an open-source, code-sharing, mindset is definitely required for progress in this field to occur.

A note could be mentioned regarding classifiers being trained for whole EEG in contrast to individual IED marking. Data scientists and research labs interested in this metric should recognize that whole EEG classification will foreseeably remain the domain of human experts due to several reasons. With the advances in automated detection comes an understanding of the limitations of algorithms and the ethics surrounding their application.[70] Hospital ethics committees or medical regulatory bodies will unlikely allow computers to make judgements on the labelling of an investigation as normal and abnormal which is to be extended without supervision to clinical care. As a parallel, most hospitals and health systems implement automated cardiac telemetry to screen for real-time diagnosis of arrhythmia. Even with the longer history of cardiac telemetry, its less complex signal characteristics and established role, human expertise and oversight is continuously needed so that unnecessary treatment is avoided. Despite this, cases have been reported of invasive interventions based on errant and artefactual automated telemetry results.[71] Governance over automated assessment versus the clinician's assessment of EEG will thus need to be closely monitored for the potential impact on treatment decisions and outcome. The focus instead should be on training, enhancing and improving the performance of IED classifiers to assist in marking and data reduction with a goal to speeding up the workflow of EEG laboratory and reviewing staff. It would be unwise to provide improved results on whole EEG classification, whereas the underlying goal of improvement desired in hospital practices is IED detection and automated marking.

Enhancing automation

A great scope of research opportunities presents itself in this field. Once a sufficiently accurate or reliable computing model for a validated detection algorithm has been developed, several other opportunities will avail themselves to enhance such models. This could incorporate future work into automated classification of the various abnormal discharges into useful subtypes. This variation can be seen, for example, in genetic generalized epilepsy or symptomatic generalized epilepsy where several kinds of epileptiform abnormalities can present themselves either between or within a single patient's EEG. The range of heterogeneity of discharges can include typical 2.5–6 Hz spike/slow wave, fragmented or localized spike or sharps, polyspike trains, polyspike/slow waves, paroxysmal fast activity[72] and also atypical rhythmic or slow spike and wave. Similarly, in focal epilepsy, one can get different morphologic, topographic and periodic characteristics, including isolated or repetitive runs (brief and long trains), which could be rhythmic or semi-rhythmic and either confined to a limited topography unilaterally or could be bilateral or multifocal. All this may be further pursued by an upgraded algorithm based on the degree of channel involvement via some quantitative criteria. Voltage topographic maps and even more advanced source localization algorithms in high-density EEG could be integrated to easily prefill quantitative sections of reports for clinicians. Future wearable devices for seizure prediction can make use of IED burden or spike rate to predict an upcoming seizure. In one study, an accuracy of 92% for seizure prediction was noted using the spike rate threshold model.[73] Predictor biomarkers currently being investigated could further allow potential predictability of pharmacoresistance in early clinical stages using large automated labelled data sets. Duration of epileptiform discharges, epileptiform burden and generalized polyspike trains, for example, are recent quantitative biomarkers associated with drug resistance.[74,75] Persyst[76] and the Encevis[64]/AIT team have been making progress in some of these domains and have commercialized their in-house AI algorithms. However, a systematic study and external validation will be required for more widespread use. This is currently being evaluated by the authors on a multi-centre data set.

Conclusion

DL algorithms, despite success in seizure detection and clinical use, have so far failed to be implemented routinely for epileptiform abnormality detection in clinical care due to inconsistent and uncertain performances. Published algorithms remain doubtful as to their generalizability and are viewed with scepticism when it comes to clinical integration in the real-world setting. Clear protocols need to be devised regarding the description of training and testing data sets utilized, annotation methods, IED-benchmarking and more thorough performance evaluation and reporting of metrics. Open sharing of source codes after model publication should be promoted to allow cross-testing and independent validation of algorithms across data sets derived from different research and hospital settings. Despite the current stumbling blocks, a new era in clinical epilepsy diagnostics with automated IED detection is likely to emerge in the near future with DL methods at the forefront.

Data availability

Data sharing is not applicable to this article as no new data were created or analysed.

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

Funding

M.J. receives support through an 'Australian Government Research Training Program (RTP) Scholarship' for PhD at Monash University, Melbourne, Australia. P.K. is supported by the Medical Research Future Fund Practitioner Fellowship (MRF1136427). L.K. is supported by the National Health and Medical Research Council (NHMRC) (GNT1183119 and GNT1160815) and the Epilepsy Foundation of America. D.N. is supported by the Graduate Research Industry Scholarship (GRIP) at Monash University, Australia. P.P. is supported by the National Health and Medical Research Council (APP1163708), the Epilepsy Foundation, The University of Melbourne, Monash University, Brain Australia and the Weary Dunlop Medical Research Foundation. T.J.O. is supported by an NHMRC Investigator Grant (APP1176426).

Competing interests

Outside the submitted work, P.P. has received speaker honoraria or consultancy fees to his institution from Chiesi, Eisai, LivaNova, Novartis, Sun Pharma, Supernus and UCB Pharma. He is an Associate Editor for Epilepsia Open. P.K.'s institution has received research grants from Biscayne Pharmaceuticals, Eisai, GW Pharmaceuticals, LivaNova, Novartis, UCB Pharma and Zynerba outside the submitted work; he has received speaker fees from Eisai, LivaNova and UCB Pharma, outside the submitted work.

Supplementary material

Supplementary material is available at *Brain Communications* online.

References

1. Gotman J, Gloor P. Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG. *Electroencephalogr Clin Neurophysiol*. 1976;41(5):513–529.
2. Frost JD. Microprocessor-based EEG spike detection and quantification. *Int J Biomed Comput*. 1979;10(5):357–373.
3. Carrie JRG. A hybrid computer technique for detecting sharp EEG transients. *Electroencephalogr Clin Neurophysiol*. 1972;33(3):336–338.
4. Saini J, Dutta M. An extensive review on development of EEG-based computer-aided diagnosis systems for epilepsy detection. *Network*. 2017;28(1):1–27.
5. Halford JJ. Computerized epileptiform transient detection in the scalp electroencephalogram: Obstacles to progress and the example of computerized ECG interpretation. *Clin Neurophysiol*. 2009;120-(11):1909–1915.
6. Abd El-Samie FE, Alotaiby TN, Khalid MI, Alshebeili SA, Aldosari SA. A review of EEG and MEG epileptic spike detection algorithms. *IEEE Access*. 2018;6:60673–60688.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521-(7553):436–444.
8. Haenssle HA, Fink C, Schneiderbauer R, *et al*. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29(8):1836–1842.
9. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402.
10. 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: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM. 2020:1–12.
11. Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: A systematic review. *J Neural Eng*. 2019;16(5):051001.
12. Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: A review. *J Neural Eng*. 2019;16(3):031001.
13. Kamitaki BK, Yum A, Lee J, *et al*. Yield of conventional and automated seizure detection methods in the epilepsy monitoring unit. *Seizure*. 2019;69:290–295.
14. Reus EEM, Visser GH, Cox FME. Using sampled visual EEG review in combination with automated detection software at the EMU. *Seizure*. 2020;80:96–99.
15. Rasheed K, Qayyum A, Qadir J, *et al*. Machine learning for predicting epileptic seizures using EEG signals: A review. *IEEE Rev Biomed Eng*. 2021;14:139–155.
16. da Silva Lourenço C, Tjepkema-Cloostermans MC, van Putten MJAM. Efficient use of clinical EEG data for deep learning in epilepsy. *Clin Neurophysiol*. 2021;132(6):1234–1240.
17. de Jong J, Cutcutache I, Page M, *et al*. Towards realizing the vision of precision medicine: AI based prediction of clinical drug response. *Brain*. 2021;144(6):1738–1750.
18. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. *Epilepsia*. 2019;60(10):2037–2047.
19. Jing J, Sun H, Kim JA, *et al*. Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. *JAMA Neurol*. 2020;77(1):103–108.
20. Brogger J, Eichele T, Aanestad E, Olberg H, Hjelland I, Aurlien H. Visual EEG reviewing times with SCORE EEG. *Clin Neurophysiol Pract*. 2018;3:59–64.
21. Moura LMVR, Shafi MM, Ng M, *et al*. Spectrogram screening of adult EEGs is sensitive and efficient. *Neurology*. 2014;83(1):56–64.
22. Wei B, Zhao X, Shi L, Xu L, Liu T, Zhang J. A deep learning framework with multi-perspective fusion for interictal epileptiform discharges detection in scalp electroencephalogram. *J Neural Eng*. 2021;18(4):0460b3.
23. Tatum WO. How not to read an EEG: Introductory statements. *Neurology*. 2013;80(1 Suppl 1):S1–S3.
24. Rathore C, Prakash S, Rana K, Makwana P. Prevalence of benign epileptiform variants from an EEG laboratory in India and frequency of their misinterpretation. *Epilepsy Res*. 2021;170:106539.
25. Benbadis SR, Kaplan PW. The dangers of over-reading an EEG. *J Clin Neurophysiol*. 2019;36(4):249.
26. Tatum WO, Selioutski O, Ochoa JG, *et al*. American Clinical Neurophysiology Society Guideline 7: Guidelines for EEG reporting. *Neurodiagn J*. 2016;56(4):285–293.
27. Moore JL, Carvalho DZ, St Louis EK, Bazil C. Sleep and epilepsy: A focused review of pathophysiology, clinical syndromes, comorbidities, and therapy. *Neurotherapeutics*. 2021;18(1):170–180.
28. Grigg-Damberger M, Foldvary-Schaefer N. Bidirectional relationships of sleep and epilepsy in adults with epilepsy. *Epilepsy Behav*. 2021;116:107735.
29. Drake ME, Pakalnis A, Phillips BB, Denio LS. Sleep and sleep deprived EEG in partial and generalized epilepsy. *Acta Neurol Belg*. 1990;90(1):11–19.
30. Seneviratne U, Lai A, Cook M, D'Souza W, Boston RC. "Sleep surge": The impact of sleep onset and offset on epileptiform

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

discharges in idiopathic generalized epilepsies. *Clin Neurophysiol*. 2020;131(5):1044–1050.
31. Dash D, Hernandez-Ronquillo L, Moien-Afshari F, Tellez-Zenteno JF. Ambulatory EEG: A cost-effective alternative to inpatient video-EEG in adult patients. *Epileptic Disord*. 2012;14(3): 290–297.
32. Seneviratne U, D'Souza WJ. Chapter 10—Ambulatory EEG. In: Levin KH, Chauvel P, eds. *Handbook of clinical neurology*, Vol. 160. *Clinical neurophysiology: Basis and technical aspects*. Elsevier; 2019:161–170.
33. Young GB, Mantia J. Continuous EEG monitoring in the intensive care unit. In: *Handbook of clinical neurology*, Vol. 140. Elsevier; 2017:107–116.
34. Louis EKS, Frey LC, Britton JW, *et al*. The normal EEG. American Epilepsy Society; 2016. Accessed 8 July 2021. https://www.ncbi.nlm.nih.gov/books/NBK390343/.
35. McKay JH, Tatum WO. Artifact mimicking ictal epileptiform activity in EEG. [Review]. *J Clin Neurophysiol*. 2019;36(4):275–288.
36. Mari-Acevedo J, Yelvington K, Tatum WO. Normal EEG variants. *Handb Clin Neurol*. 2019;160:143–160.
37. Kang JY, Krauss GL. Normal variants are commonly overread as interictal epileptiform abnormalities. [Review]. *J Clin Neurophysiol*. 2019;36(4):257–263.
38. Nayak CS, Anilkumar AC. EEG normal sleep. In: *StatPearls* [Internet]. StatPearls Publishing, 2021. Available from: https://www.ncbi.nlm.nih.gov/books/NBK537023/.
39. Nhu D, Janmohamed M, Perucca P, *et al*. Graph convolutional network for generalized epileptiform abnormality detection on EEG. In: IEEE Signal Process Med Biol Symp SPMB. 2021.
40. Lin L, Drislane FW. Lateralized periodic discharges: A literature review. *J Clin Neurophysiol*. 2018;35(3):189–198.
41. Meritam Larsen P, Wüstenhagen S, Terney D, *et al*. Photoparoxysmal response and its characteristics in a large EEG database using the SCORE system. *Clin Neurophysiol*. 2021; 132(2):365–371.
42. Buluş E, Abanoz Y, Gülen Abanoz Y, Yeni SN. The effect of cognitive tasks during electroencephalography recording in patients with reflex seizures. *Clin EEG Neurosci*. 2022;53(1):54–60.
43. Gelžinienė G, Endzinienė M, Jurkevičienė G. EEG activation by neuropsychological tasks in idiopathic generalized epilepsy of adolescence. *Brain Dev*. 2015;37(4):409–417.
44. Webber WRS, Litt B, Lesser RP, Fisher RS, Bankman I. Automatic EEG spike detection: What should the computer imitate? *Electroencephalogr Clin Neurophysiol*. 1993;87(6):364–373.
45. Wilson SB, Harner RN, Duffy FH, Tharp BR, Nuwer MR, Sperling MR. Spike detection. I. Correlation and reliability of human experts. *Electroencephalogr Clin Neurophysiol*. 1996;98(3): 186–198.
46. Bagheri E, Dauwels J, Dean BC, Waters CG, Westover MB, Halford JJ. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clin Neurophysiol*. 2017;128(10): 1994–2005.
47. Kural MA, Duez L, Sejer Hansen V, *et al*. Criteria for defining interictal epileptiform discharges in EEG: A clinical validation study. *Neurology*. 2020;94(20):e2139–e2147.
48. Beniczky S, Aurlien H, Franceschetti S, *et al*. Interrater agreement of classification of photoparoxysmal electroencephalographic response. *Epilepsia*. 2020;61(9):e124–e128.
49. Piccinelli P, Viri M, Zucca C, *et al*. Inter-rater reliability of the EEG reading in patients with childhood idiopathic epilepsy. *Epilepsy Res*. 2005;66(1):195–198.
50. Halford JJ, Arain A, Kalamangalam GP, *et al*. Characteristics of EEG interpreters associated with higher interrater agreement. *J Clin Neurophysiol*. 2017;34(2):168–173.
51. Golmohammadi M, Harati Nejad Torbati AH, Lopez de Diego S, Obeid I, Picone J. Automatic analysis of EEGs using big data and hybrid deep learning architectures. *Front Hum Neurosci*. 2019;13:76.
52. Seneviratne U, Hepworth G, Cook M, D'Souza W. Atypical EEG abnormalities in genetic generalized epilepsies. *Clin Neurophysiol*. 2016;127(1):214–220.
53. Obeid I, Picone J. The Temple University hospital EEG data corpus. *Front Neurosci*. 2016;10:196.
54. 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):1174–1179.
55. Lourenço C, Tjepkema-Cloostermans MC, Teixeira LF, van Putten MJAM. Deep learning for interictal epileptiform discharge detection from scalp EEG recordings. In: Henriques J, Neves N, de Carvalho P, editors. *IFMBE Proceedings*, Vol 76. Springer, 2020:1984–1997.
56. Prasanth T, Thomas J, Yuvaraj R, *et al*. Deep learning for interictal epileptiform spike detection from scalp EEG frequency sub bands. *Annu Int Conf IEEE Eng Med Biol Soc*. 2020;2020:3703–3706.
57. Thomas J, Thangavel P, Peh WY, *et al*. Automated adult epilepsy diagnostic tool based on interictal scalp electroencephalogram characteristics: A six-center study. *Int J Neural Syst*. 2021;31(5): 2050074.
58. Fukumori K. Fully data-driven convolutional filters with deep learning models for epileptic spike detection. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings. 2019:2772–2776.
59. Thomas J, Comoretto L, Jin J, Dauwels J, Cash SS, Westover MB. EEG CLassification Via convolutional neural network-based interictal epileptiform event detection. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018:3148–3151.
60. Xu Z, Wang T, Cao J, Bao Z, Jiang T, Gao F. BECT spike detection based on novel EEG sequence features and LSTM algorithms. *IEEE Trans Neural Syst Rehabil Eng*. 2021;29:1734–1743.
61. Lashgari E, Liang D, Maoz U. Data augmentation for deep-learning-based electroencephalography. *J Neurosci Methods*. 2020;346:108885.
62. Koren J, Hafner S, Feigl M, Baumgartner C. Systematic analysis and comparison of commercial seizure-detection software. *Epilepsia*. 2021;62(2):426–438.
63. Jaramillo M. Persyst: The worldwide leader in EEG software. Persyst. Accessed 8 October 2021. https://www.persyst.com/.
64. Spike detection—Encevis. encevis. Accessed 3 October 2021. https://www.encevis.com/solutions/spike-detection/.
65. Scheuer M. Spike detection: Inter-reader agreement and a statistical Turing test on a large data set. *Clin Neurophysiol* 2017;128(1): 243–250.
66. Halford JJ, Westover MB, LaRoche SM, *et al*. Interictal epileptiform discharge detection in EEG in different practice settings. *J Clin Neurophysiol*. 2018;35(5):375–380.
67. Reus EEM, Cox FME, van Dijk JG, Visser GH. Automated spike detection: Which software package? *Seizure*. 2022;95:33–37.
68. Raghu S, Sriraam N, Gommer ED, *et al*. Cross-database evaluation of EEG based epileptic seizures detection driven by adaptive median feature baseline correction. *Clin Neurophysiol*. 2020;131(7): 1567–1578.
69. Bernabei JM, Owoputi O, Small SD, *et al*. A full-stack application for detecting seizures and reducing data during continuous electroencephalogram monitoring. *Crit Care Explor*. 2021;3(7):e0476.
70. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. 2020;46(3):205–211.
71. Henriques-Forsythe MN, Ivonye CC, Jamched U, Kamuguisha LKK, Olejeme KA, Onwuanyi AE. Is telemetry overused? Is it as helpful as thought? *Cleve Clin J Med*. 2009;76(6):368–372.
72. Sagi V, Kim I, Bhatt AB, Sonmezturk H, Abou-Khalil BW, Arain AM. Generalized paroxysmal fast activity in EEG: An unrecognized finding in genetic generalized epilepsy. *Epilepsy Behav*. 2017;76: 101–104.

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022

73. Slimen IB, Boubchir L, Seddik H, *et al.* Epileptic seizure prediction based on EEG spikes detection of ictal-preictal states. *J Biomed Res*. 2020;34(3):162.
74. Sun Y, Seneviratne U, Perucca P, *et al.* Generalized polyspike train: An EEG biomarker of drug-resistant idiopathic generalized epilepsy. *Neurology*. 2018;91(19):e1822–e1830.
75. Arntsen V, Sand T, Syvertsen MR, Brodtkorb E. Prolonged epileptiform EEG runs are associated with persistent seizures in juvenile myoclonic epilepsy. *Epilepsy Res*. 2017;134: 26–32.
76. Spike Detection—Persyst. Persyst. Accessed 3 October 2021. https://www.persyst.com/technology/spike-detection/.

Downloaded from https://academic.oup.com/braincomms/article/4/5/fcac218/6678128 by University of Melbourne Library user on 21 October 2022