EBioMedicine 000 (2021) 103275

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data

Subhrajit Roy[a,1,2], Isabell Kiral[a,1,3], Mahtab Mirmomeni[a,1], Todd Mummert[b], Alan Braz[b], Jason Tsay[b], Jianbin Tang[a], Umar Asif[a], Thomas Schaffter[h], Mehmet Eren Ahsen[j,5], Toshiya Iwamori[d], Hiroki Yanagisawa[d], Hasan Poonawala[e,4], Piyush Madan[f], Yong Qin[g], Joseph Picone[i], Iyad Obeid[i], Bruno De Assis Marques[a], Stefan Maetschke[a], Rania Khalaf[f], Michal Rosen-Zvi[c], Gustavo Stolovitzky[b,*], Stefan Harrer[a,6,*], IBM Epilepsy Consortium

[a] IBM Research, Melbourne, Australia
[b] IBM Research, T. J. Watson Research Center, USA
[c] IBM Research, Haifa, Israel
[d] IBM Research, Tokyo, Japan
[e] IBM Research, Singapore
[f] IBM-MIT AI Lab, Cambridge, USA
[g] IBM Research, Beijing, China
[h] Sage Bionetworks, Seattle, USA
[i] College of Engineering, Temple University, Philadelphia, USA
[j] Icahn School of Medicine at Mount Sinai, NY, USA

ARTICLE INFO

Article History:
Received 30 November 2020
Revised 21 February 2021
Accepted 23 February 2021
Available online xxx

Keywords:
Epilepsy
Seizure detection
Artificial intelligence
Deep neural networks
EEG
Automatic labelling, Crowdsourcing challenges

ABSTRACT

Background: Assistive automatic seizure detection can empower human annotators to shorten patient monitoring data review times. We present a proof-of-concept for a seizure detection system that is sensitive, automated, patient-specific, and tunable to maximise sensitivity while minimizing human annotation times. The system uses custom data preparation methods, deep learning analytics and electroencephalography (EEG) data.

Methods: Scalp EEG data of 365 patients containing 171,745 s ictal and 2,185,864 s interictal samples obtained from clinical monitoring systems were analysed as part of a crowdsourced artificial intelligence (AI) challenge. Participants were tasked to develop an ictal/interictal classifier with high sensitivity and low false alarm rates. We built a challenge platform that prevented participants from downloading or directly accessing the data while allowing crowdsourced model development.

Findings: The automatic detection system achieved tunable sensitivities between 75.00% and 91.60% allowing a reduction in the amount of raw EEG data to be reviewed by a human annotator by factors between 142x, and 22x respectively. The algorithm enables instantaneous reviewer-managed optimization of the balance between sensitivity and the amount of raw EEG data to be reviewed.

Interpretation: This study demonstrates the utility of deep learning for patient-specific seizure detection in EEG data. Furthermore, deep learning in combination with a human reviewer can provide the basis for an assistive data labelling system lowering the time of manual review while maintaining human expert annotation performance.

Funding: IBM employed all IBM Research authors. Temple University employed all Temple University authors. The Icahn School of Medicine at Mount Sinai employed Eren Ahsen. The corresponding authors

The IBM Epilepsy Consortium has the following members: Sharon Hensley Alford, Rachita Chandra, Wen Liu, Wei Lian Ti, Li Ma, Michael Cherner, Dario Arcos-Diaz, Paul Hake

* Corresponding authors.
E-mail addresses: gustavo@us.ibm.com (G. Stolovitzky), stefan.harrer@dhcrc.com (S. Harrer).
[1] These authors contributed equally to this work.
[2] Google Brain, London, UK
[3] Amazon Web Services, Melbourne, AU
[4] Amazon Web Services, London, UK
[5] Department of Business Administration, University of Illinois at Urbana-Champaign, Champaign, IL, USA
[6] Digital Health Cooperative Research Centre, Melbourne, AU

https://doi.org/10.1016/j.ebiom.2021.103275

Please cite this article as: S. Roy et al., Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data, EBioMedicine (2021), https://doi.org/10.1016/j.ebiom.2021.103275

Stefan Harrer and Gustavo Stolovitzky declare that they had full access to all the data in the study and that they had final responsibility for the decision to submit for publication.

Research in context

Evidence before this study

Epilepsy is a highly individualized neurological condition with disease expressions changing over time. It thus cannot be diagnosed, treated and managed in a uniform and equally efficient way across patients. The ability to monitor patients individually and continuously and to log seizure episodes in disease diaries is key to gaining a patient-specific understanding of the disease. This can empower doctors to optimize and adjust medication response and pharmaceutical companies to design more efficient clinical trials. Until recently such disease diaries were created manually making data review a highly cost- and time-intensive process and rendering records inaccurate when populated by patients through self-reporting outside the clinic.

Added values of this study

New deep learning-based techniques allow automatic detection of seizures in brain activity data monitored by electroencephalography (EEG) sensors. Harnessing the power of crowdsourced artificial intelligence algorithm development and one of the largest EEG datasets in existence we have demonstrated an automatic seizure detection model which can assist human reviewers in substantially reducing the amount of raw EEG data to be annotated manually.

Implications of all the available evidence

Our system uses deep learning technology and custom data preparation techniques to automatically learn patient-specific seizure signatures. It then filters seizure segments out of raw EEG data for verification by an expert neurologist. Our system can be tuned by the human reviewer to achieve detection sensitivities in excess of 90% or raw data reduction factors of up to 142x.

1. Introduction

This decade has seen an ever-growing number of scientific fields benefitting from the advances in machine learning technology and tooling. More recently, this trend reached the medical domain [1], with applications ranging from cancer diagnosis [2,3], prediction of acute kidney injury [4], detection of diabetic retinopathy [5], mining of electronic health records [6] to brain-machine-interfaces [7,8]. While Kaggle has pioneered the crowdsourcing of machine learning challenges to incentivize data scientists from around the world to advance algorithm and model design, the increasing complexity of healthcare domain problems demands interdisciplinary teams with expertise in data science, the problem domain, and competent software engineers with access to large compute resources. Teams or people who meet these criteria are few and far between, leading to a small pool of possible participants and a loss of experts dedicating their time to solving important problems. Participation is even further restricted in the context of any challenge run on confidential use cases or with sensitive data.

In order to protect such sensitive and proprietary data, while at the same time enabling a crowdsourced challenge, we have recently introduced a challenge ecosystem that utilizes the so-called model-to-data paradigm pioneered by the DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenges [9,10]. This approach allows the solver community to submit their models to the platform which will then autonomously organize model training and testing in a secure cloud environment and provide feedback on model performance to participants. Solvers can then use the model performance to improve their algorithms. In this scheme, the participants cannot download or directly access the challenge data at any point but have the full suite of crowdsourced challenge tools at their disposal. This challenge concept opens the door to running crowdsourced challenges and to enabling broad public benchmarking against proprietary or sensitive datasets which cannot be made publicly available [11]. Using this idea, we recently designed and ran the Deep Learning Epilepsy Detection Challenge to crowdsource the development of an automated labelling system for brain recordings, aiming to advance epilepsy research.

Epilepsy is a neurological disease that affects over 1% of the world population [12]. Patients suffer from sudden and unexpected seizures which impact their physical health and mental wellbeing [13]. Being a highly individualized condition, its expression changes from patient to patient. Even a specific patient's pathology can vary over time. This makes adequate diagnosis, treatment, and disease management extremely challenging: one third of all epilepsy patients suffer from refractory epilepsy. Two-thirds of patients respond to medication in some way at some point in their journey, but oftentimes the little understood evolving nature of the disease leads to fading or transient therapeutic control [12].

The most common method of tackling this challenge is to monitor patients continuously and log disease episodes of relevance in disease diaries [14]. These longitudinal data repositories can then be used to investigate and adjust the effect of medication in quasi-real-time, and to study the correlation between treatment regimens and disease progression. While this data-driven approach to treatment management and *in-situ* care optimization is seen as key to fundamentally changing the success of treatment and efficiency of clinical trials [15] until recently, real-world implementations of disease diaries have been entirely manual and thus highly inefficient. Manually created disease diaries are only approximately 50% accurate [16]. This is not rooted in sloppy reporting techniques. It is the individualized and incapacitating nature of the disease itself that leaves patients unable to recognize, remember, or keep track of their own seizures. That makes it impossible for untrained observers to recognize and describe seizure episodes in clinically actionable ways [15]. In order to overcome this challenge, and to leverage a plethora of wearable and mobile sensing platforms, the field has turned to exploring the use of machine learning techniques for the development of automatic patient monitoring systems [17].

Amongst a broad spectrum of sensor modalities ranging from video cameras to smart watches [18], the electroencephalogram (EEG), which uses scalp electrodes, is considered to be the gold standard for seizure monitoring in clinical as well as non-clinical environments [13]. However, while EEG monitoring systems have evolved from relying on intracranial implanted electrodes to use of non-invasive clinical and non-clinical wearable devices, automatic annotation of EEG data remains a challenging machine learning problem. Primary reasons for this include low signal to noise ratio, movement artefacts, poor electrical conduction and nonlinearly distorted

Please cite this article as: S. Roy et al., Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data, EBioMedicine (2021), https://doi.org/10.1016/j.ebiom.2021.103275

crosstalk between spatially adjacent sensors. Disease-specific intricacies such as the highly individualized profiles of seizure patterns make generalizability of detection models across patients challenging. As a result, in today's practice, EEGs are still interpreted manually, or 'read' by trained neurologists. The associated time and cost burdens are substantial and account for approximately 5% of the total hospital charges for epilepsy patients admitted to Intensive Care Units (ICUs) in the US [13]. Furthermore, doctors responsible for this highly repetitive and time-consuming process find themselves caught between the equally undesirable options of either having to limit the time they can devote to attend to their patients [19], reducing the duration of monitoring sessions or reducing the amount of data to be manually reviewed [13].

A variety of machine learning (ML)-based automatic EEG annotation systems have been proposed [13,20] to reduce this burden. Of special interest are deep-learning models as they can learn to automatically recognise different seizure patterns for individual patients which allows to calibrate these detection algorithms to patient-specific disease expressions. Some have been deployed and tested in clinical scenarios [21,22]. Clinical acceptance of this technology has been slow [23]. The lack of commonly adopted performance metrics to evaluate performance and compare to human expert reviewers [24] has inhibited broad adoption of these systems in critical care settings. Generalisability of performance across datasets collected at different institutions has been problematic also [13].

Using one of the world's largest EEG datasets, the TUH Seizure Corpus [25,26], the Deep Learning Epilepsy Detection Challenge tasked participants to develop deep learning models for automatic annotation of epileptic seizure signals in raw EEG data with maximum sensitivity and minimum false alarm rates. Using the Time-Aligned Event Scoring (TAES) metric, an evaluation framework custom-designed to score high-resolution automatic EEG annotation algorithms [24], we assessed the potential of these annotation models for use by clinical neurologists as assistive labelling systems for raw EEG monitoring data.

In the following sections we describe the architecture and functionality of our custom-developed crowdsourcing challenge platform, with a special focus on its model-to-data feature, the design and execution of the Deep Learning Epilepsy Detection Challenge, as well as the scientific outcomes and validation results of the best performing participant models.

2. Methods

With a goal to run a challenge that mobilizes the largest possible pool of participants globally across IBM, we designed a crowdsourced challenge called the Deep Learning Epilepsy Detection Challenge. Participants were asked to develop an automatic labelling system to reduce the time a clinician would need to diagnose patients with epilepsy. Labelled data for the challenge were provided by Temple University Hospital (TUH) [22,26]. We partitioned this data to create training, validation and blind test sets which participants could work with only through our platform.

To provide an experience with a low barrier of entry, and to demonstrate that following the model-to-data paradigm a crowdsourced challenge can run efficiently without participants ever having to directly access or download the challenge data, we designed a generalizable challenge platform based on the following principles: (1) eliminate the need of in-depth knowledge of the specific domain. (i.e. no participant should need to be a neuroscientist or epileptologist); (2) eliminate the need of more than basic programming knowledge (i.e. no participant should need to learn how to process fringe data formats and stream data efficiently), (3) eliminate the need for participants to provide their own computing resources, and (4) eliminate the need for participants to download or directly access the challenge data in any way.

The platform guided participants through the entire process from sign-up to model submission, facilitated collaboration, and provided instant feedback to the participants through data visualization and intermediate online leaderboards. The competitive phase of the Deep Learning Epilepsy Detection Challenge ran for 6 months. Twenty-five teams, with a total number of 87 data scientists and software engineers from 14 global IBM locations participated. Seven teams submitted final solutions five of which were valid final submissions as per the challenge rules.

2.1. Study design

2.1.1. The Deep Learning Epilepsy Detection Challenge platform

The architecture of the platform that was designed and developed as well as data and model flow through it during the challenge are shown in Fig. 1. The entire system consists of a number of interacting components:

(1) A web portal serves as the entry point to challenge participation, providing challenge information, such as timelines and challenge rules, scientific background information and a description of the data used for this challenge. The portal also facilitated the formation of teams and provided participants with an intermediate leaderboard of submitted results and a final leaderboard at the end of the challenge. A screenshot of the starting page of the web portal can be found in the supplemental information (Fig. S13). (2) IBM Watson Studio [27] is the umbrella term for a number of services offered by IBM and accessible to participants. Upon creation of a user account through the web portal, an IBM Watson Studio account was automatically created for each participant that gave users access to the (3) IBM Data Science Experience (DSX) platform which hosted a user interface and starter kit and formed the main component for designing and testing models during the challenge. DSX allows for real-time collaboration on shared notebooks between team members. A starter kit in the form of Jupyter notebooks [28], supporting the popular deep learning libraries TensorFLow [29] and PyTorch [30], was provided to all teams to guide them through the challenge process. Upon instantiation, the starter kit loaded the necessary python libraries and custom functions for the invisible integration with (4) IBM Cloud Object Storage (COS) [31] and the analytics engine (5) Watson Machine Learning (WML). In dedicated notebook cells, participants could develop custom pre-processing code (including custom montages), machine learning models, and post-processing algorithms. The starter kit provided instant feedback about participants' custom routines through data visualizations. Using the notebook only, teams were able to run their code on WML, making use of a compute cluster of IBM's resources. The starter kit also enabled submission of the final code to a data storage to which only the challenge team had access. WML provided access to shared compute resources (Graphics Processing Units, GPUs). Code was bundled automatically into the starter kit and deployed on WML. WML in turn had access to shared storage from which it requested recorded data and to which it stored the participant's code and trained models. The data for this challenge resided in COS. Note that using the starter kit, participants submitted their model code to the platform which autonomously organized model training and validation on the raw data and provided back model performance results to participants. The participants could then investigate this feedback in order to better design custom algorithms. This approach is called a model-to-data paradigm which unlike in Kaggle-style challenge scenarios keeps data shielded from the solver community while at the same time allowing a crowdsourced approach to model development. (6) Utility Functions were loaded into the starter kit at instantiation. This set of functions included code to pre- and post-process data into a more common format while preserving all seizure related information, to optimize streaming through the use of the NutsFlow and NutsML libraries [32], and to provide seamless access to all services

Please cite this article as: S. Roy et al., Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data, EBioMedicine (2021), https://doi.org/10.1016/j.ebiom.2021.103275

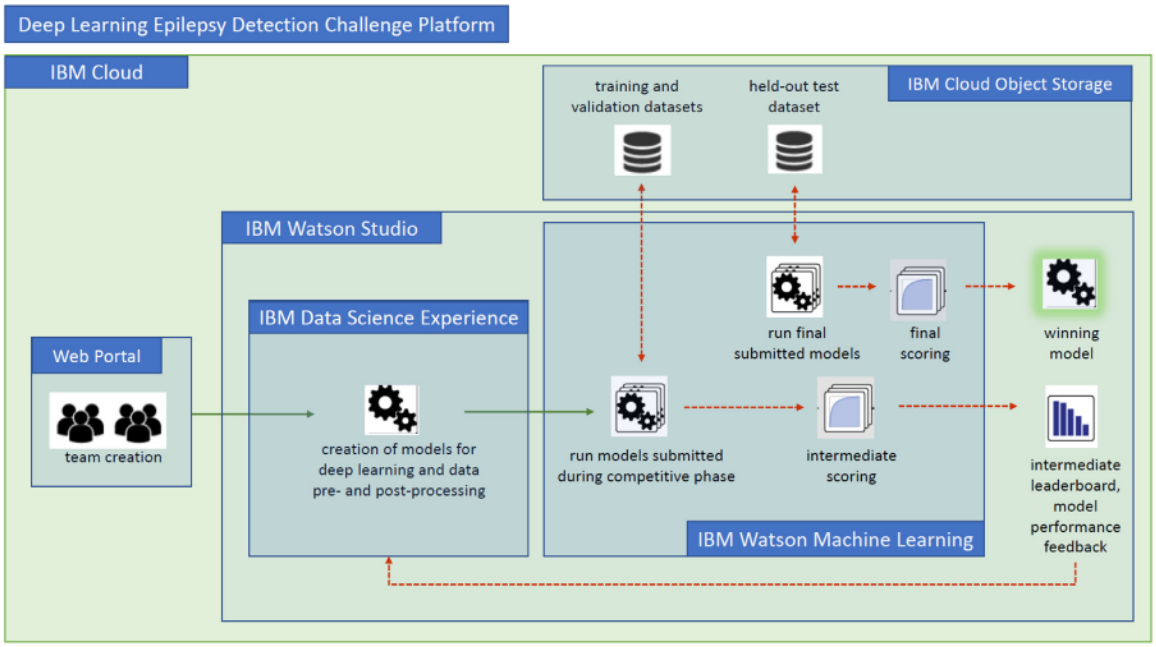

Fig. 1. A block diagram of the high-level architecture of the custom-built challenge platform that depicts data and model flow during challenge operation. In this model-to-data paradigm challenge participants at no point download or access the data directly. Instead they create and submit models to the platform (green solid arrows) which automatically organises training and testing and then provides feedback on model performance to participants (orange dashed arrows). This is fundamentally different to conventional crowdsourced challenge setups. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

used. (7) Final code scoring after completion of the challenge was conducted in an automated way as soon as code was submitted through the starter kit.

2.1.2. Data sources and preparation

All data used in this study is available as open source data at the web site: https://www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml. This data was collected at Temple University Hospital, a research and teaching hospital. The study has been conducted under the approval of the Temple University Institutional Review Board (IRB) under IRB No. 20,774, which supports the release of data once it has been properly anonymized. The IRB has been in existence since 2012 and has been renewed on an annual basis (currently through 2021). Patients consent to the use of their data for research and teaching through a written consent as part of their admission record at Temple Hospital. The data has been carefully anonymized before being released from Temple Hospital so that a patient's identity cannot be reconstructed from the data. Files mapping the anonymized data to identifiable data are maintained with Temple Hospital's Health Insurance Portability and Accountability Act (HIPAA)-protected network and never leave the hospital. The study has been conducted in compliance with this ethical approval.

The TUH EEG Seizure Corpus v1.2.0 [22] which contains scalp EEG records of 315 patients with annotated seizure times was split into training and validation datasets for the challenge (Table 1). The dataset is composed of 822 monitoring sessions with 280 sessions containing a total of 2012 seizures. Annotation protocols including explanations of data collection and split processes are provided on the TUH Open Source EEG Resources platform [25]. The validation dataset was used to determine team rankings on the intermediate leaderboard during the competitive phase (Fig. S1, supplemental information). Another dataset containing annotated data from 50 patients following the same format as v1.2.0 was used as a blind held-out test dataset (Table 1) for final team rankings on the final leaderboard at the end of the challenge (Fig. S1, supplemental information). After completion of the challenge this blind test dataset was merged with v1.2.0 and made publicly available as version v1.2.1 of the TUH seizure corpus thus allowing reproducibility of and continuous benchmarking against the results published in this paper.

The size of training, validation, and blind test sets are shown in Table 1. Training and validation datasets were composed to reflect a balanced demographic profile (49.5% of patients in the training dataset are male, 44% of patients in the validation dataset are male, further demographic distributions for the datasets are provided in [22]). Training and validation sets were used during the competitive phase of the challenge following the model-to-data paradigm described above not allowing participants any direct access to or downloading of any of the data. The blind, held-out test set was not accessible to participants' models at any time during the challenge and was only used once by the challenge organising team to evaluate the submitted models during the scoring phase after the completion of the competitive phase (see Fig. S1 supplemental information for challenge timeline).

The TUH EEG Seizure Corpus consists of EEG sessions recorded according to the 10/20 electrode configuration [33] and utilizing the European Data Format (EDF) [26]. We converted the recorded EEG signal into a set of montages, or differentials, of electrode signals based on guidelines proposed by the American Clinical Neurophysiology Society [34]. In this challenge, we used the transverse central parietal (TCP) montage system for accentuating spike activity which has been shown to improve performance in EEG classification tasks [35].

Table 1
Number and types of samples in training, validation and blind test sets. Detailed demographic distributions are provided in [22].

	Training set	Validation set	Blind test set
Patients	265	50	50
EDF files	2032	1032	1022
Seizure [s]	76,517	55,764	39,464
Non-seizure [s]	1,119,863	562,331	503,670
Total [s]	1,196,381	618,096	543,134

2.1.3. Evaluation procedure

The evaluation of machine learning algorithms for seizure detection lacks standardization. Typically, two different types of methods are used: epoch-based and term-based. Epoch-based methods compute a summary score decision per unit of time. Term-based methods score on an event basis and do not count individual frames.

Both methods have disadvantages. While epoch-based scoring generally weighs duration of events more heavily, term-based methods are a permissive way of scoring and can result in artificially high sensitivities. In this challenge, we use a method called Time-Aligned Event Scoring (TAES) that utilizes concepts of both epoch-based and term-based methods. It considers percentage overlap between reference and hypothesis and weighs errors accordingly. The TAES metric is described in detail in [24]. Note that since TAES weighs both the number and duration of identified seizures, the sensitivity vs. false positive profile is not the same as for standard methods where sensitivity typically increases with an increasing false positive rate. In TAES the sensitivity is penalized at both low and high false positives. For low false alarms the sensitivity is low since enough seizures are not being discovered by the classifier. At high false alarm rates, since most samples are marked as seizures, although the total duration of identified seizures is high, the number of unique seizures identified is low and thus TAES again penalizes the sensitivity value.

Evaluation Metric: The two qualities of an automatic seizure detection system should be high sensitivity and low false alarm rate. For the purpose of this challenge, we use the following metric to combine these two parameters into an evaluation metric E with $E = (FA / S) - \varepsilon * S$ where FA is False Alarm per 24 h, S is Sensitivity, and ε is a positive constant. The best solution will have the smallest E. Note that E has two contributing terms. The first term FA/S ensures that systems with lower FA and higher S are preferred. The second term ensures that higher S solutions are preferred if for two systems the (FA/S) ratio is same. This formula constitutes the pre-defined objective function for measuring success and remained unchanged during the course of this challenge.

Scoring: During the competitive phase of the challenge scoring happened instantaneously: Once a model had been trained, it was evaluated using a validation data set and the score was submitted, displayed and ranked against other participants' models in the leaderboard section of the challenge portal. During the evaluation phase (i.e. after completion of the competitive phase) we gave participants a 2-week time window to submit their final trained model. We extracted the pre-processing model and post-processing code from each submission and ran these models on a held-out blind test dataset (to which participants had not had access to at any point during the challenge). This was the final submission evaluation similar to the "private leaderboard" in Kaggle. In Kaggle, this "private leaderboard" is also immediate since one submits only the predictions. For our challenge, we ran the participants' final submitted code on the blind test dataset, which took 3 weeks to complete for all final submissions. The reason for deviating from conventional Kaggle-style protocol by submitting only predictions is that unlike Kaggle we keep raw data confidential and do not provide it to participants at any point.

2.1.4. Role of funding source

The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

3. Results

At the completion of the challenge, 7 teams submitted their final algorithms which were evaluated against the blind test set. Upon review of all final submissions, we found that 5 out of the 7 teams had made valid submissions as per the challenge rules. These 5 teams were named Ids_cpmp, Otameshi, AI4MH, Team SG, and EpiInsights. They were considered in the final evaluation stage. Several measures of performance obtained using the validation dataset (leaderboard) and the blind test set are provided in Fig. 2: evaluation metric (E), sensitivity (S), false alarm rate (FA/24 h) as well as a sensitivity vs. FA/24 h plot for all 5 submissions. It can be seen that the performance of the 5 submissions is similar in both validation and test sets, indicating that there is no evidence of overfitting.

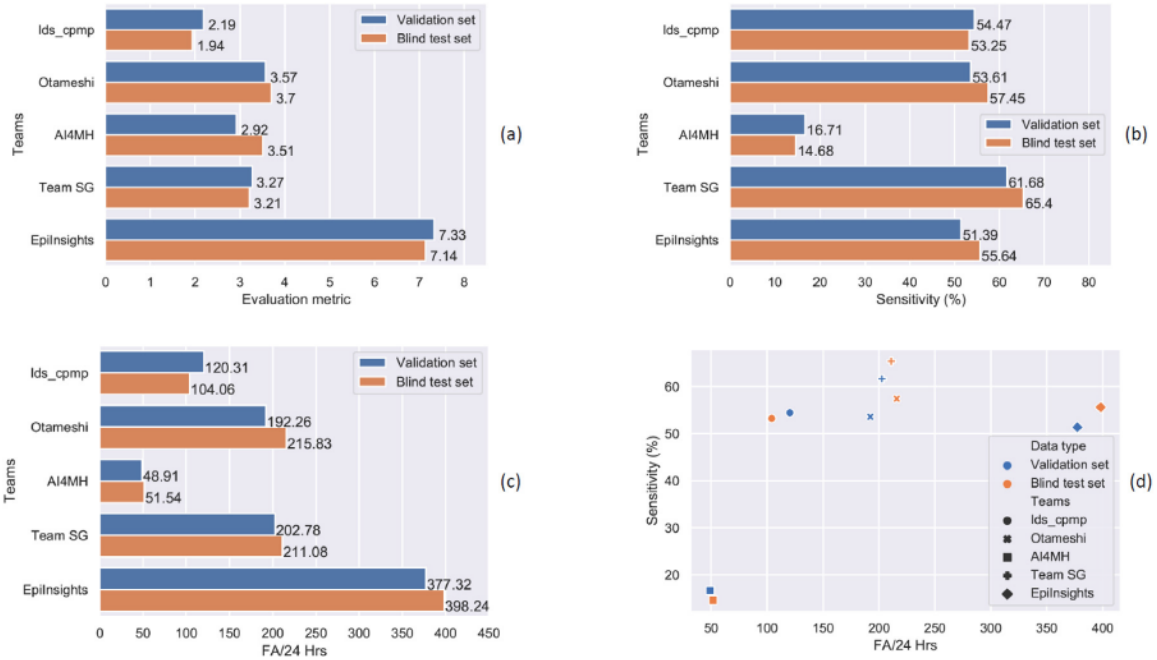

Fig. 2. All 5 valid final submissions were tested against validation and blind test sets. The plots show the results for (a) evaluation metric (E), (b) sensitivity (S), (c) false alarm rate (FA/24 h and (d) sensitivity (S) plotted as a function of FA/24 h.

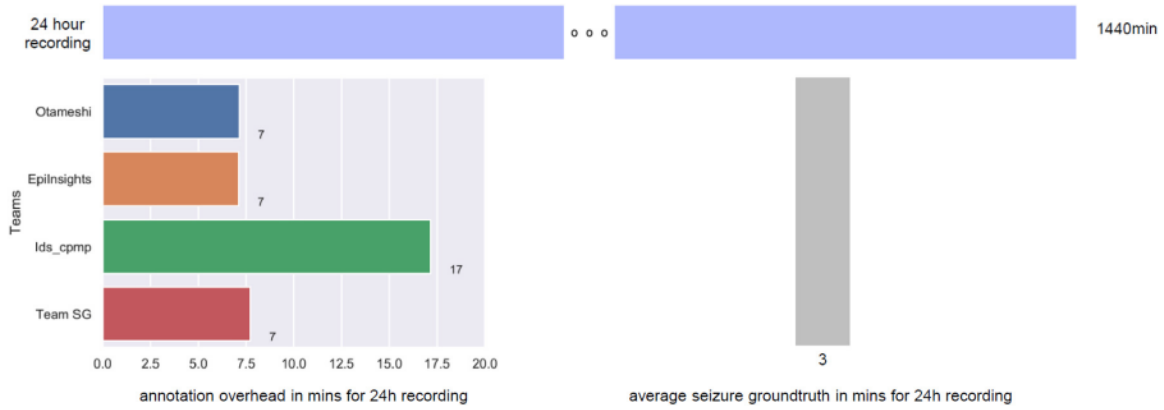

Fig. 3. In order to label 24 h of EEG recordings an unassisted human annotator has to review all 24 h of raw EEG data (top). Using the systems developed in this challenge, the amount of data needing review is the sum of the seizure ground truth (correctly detected true positive actual seizure segments) plus the annotation overhead (incorrectly detected false positive segments). All 4 automatic systems operate at 75% detection sensitivity. A conservative upper bound approximation for the total seizure ground truth duration in a 24 h raw EEG data recording is ~0.2% [22] or ~3 min. The best models achieve a minimum annotation overhead of 7 min which therefore allows to reduce the total amount of raw EEG data to be reviewed by a human annotator from 24 h down to 10 min or less. Note that the duration of seizure ground truth may fluctuate across patients, i.e. a patient might experience longer or more frequent seizure episodes on certain days which impacts the total duration of raw EEG data to be reviewed for that day. The annotation overhead however remains unaffected and will stay at the levels shown in the figure for all patients at all times.

Any automatic seizure detection system, be it a retrospective assistive labelling system or a real-time alert system, needs to be at least as sensitive as a human observer for it to be clinically relevant. This sensitivity goal for an automated system is 75% [24,36]. For false alarm rates equal to or lower than those of human observers the detection system could replace monitoring clinicians. For false alarm rates higher than those of human observers the system is not suitable to replace them, but for low enough false alarm rates such a system can be used as data reduction tool which decreases the amount of raw EEG data a human annotator needs to review. While unassisted human annotators will review the entirety of all raw EEG data, use of an assistive labelling system allows review of only those EEG segments which the system detects: both, correctly in terms of true positives (actual ictal segments) and incorrectly in form of false positives (false alarms, actual non-ictal segments). We call the total amount of raw EEG data composed by all accumulated false positive segments the annotation overhead, and the total duration of raw EEG data defined by all true positives the annotation ground truth. In the following section we show that 4 out of the 5 automatic seizure detection systems developed in this challenge could be used to reduce the annotation overhead by up to several orders of magnitude thus substantially decreasing the labelling time burden for human annotators.

At their lowest false alarm rate levels none of the 5 final submission models reached 75% detection sensitivity thus rendering the developed algorithms unsuitable as real-time alert systems (Fig. 2 (b)). However, as part of their final solution, team Otameshi and Ids_cpmp introduced an engineering post-processing step which added synthetic false alarms (details provided in supplemental information). This step introduced a hyperparameter which allowed for the tuning of sensitivity and FA rate of the developed models. We removed this engineering step for producing the results shown in Fig. 2 to be able to assess detection performance at the lowest achievable false alarm rates for all submissions. We then added the engineering step back into the final submissions of all 5 teams which allowed an increase in sensitivity to above the 75% level threshold for all submissions except for the one from team AI4MH which is therefore excluded from the following analyses (detailed reproducible individual descriptions of all solutions including neural network architectures, hyperparameter selection procedures, training and scoring methods, data pre- and post-processing techniques and visual solution flow charts are provided in section III of the supplemental information). The total false alarm numbers per 24 h obtained by each of these four submissions at 75% sensitivity are shown in Fig. 4a and yield the shortest achievable annotation overheads for each automatic seizure detection system as depicted in Fig. 3.

Teams Otameshi, EpiInsights and Team SG all achieve minimum annotation overheads of 7 min. In good approximation it can be assumed that on average ~0.2% or ~3 min of a continuous 24h-long raw EEG recording describe ictal segments while 98.8% or 1437 min of the raw data are correlated with non-ictal episodes [22]. For unassisted human labelling of 24 h of raw EEG data this means that the seizure ground truth is 3 min and the annotation overhead is 1437 min. Using the automatic labelling systems reduces the annotation overhead to 7 min thus reducing the amount of total raw EEG data that needs to be reviewed by a human expert from 24 h to 10 min.

Note that we do not claim this time to be the time that it would take a human annotator to label the data. Actual human annotation times are determined by annotation procedures, review protocols as well as the degree of expertise and practice of the reviewers. Regardless of these factors the assistive detection systems described above reduce the overall amount of data that needs to be reviewed by up to two orders of magnitude with a maximum achievable reduction factor of 142x (Fig. 4b) and thus lead to a substantial decrease of the time and cost burden for all human annotation scenarios.

Further investigating the effect of the engineering step introduced by teams Otameshi and Ids_cpmp, we found that as more false alarms are included the sensitivity reaches a maximum and then decreases again. This effect can be attributed to the impact of the TAES evaluation metric which penalizes both low and high false alarms as explained above. Fig. 5 plots the path from 75% sensitivity to maximum achievable sensitivity against false alarm rates for all 4 submissions. With increasing false alarm rates, the respective data reduction factors decrease (Fig. 6). Exploiting this effect allows the development of a tunable assistive labelling system: annotation sensitivities beyond 90% can be achieved but come at the cost of lower data reduction factors, i.e. the price for higher labelling sensitivity is longer data review time. This tunability allows clinical experts to cater the quality of their annotation services to healthcare provider and insurer specific frameworks: depending on the amount of billable time for data review and the amount of data to be reviewed, a custom data reduction factor can be calculated that compresses the total raw data to the exact size that can be reviewed during the billable time while at the same time optimizing annotation sensitivity.

Note that three systems (Otameshi, EpiInsights and Team SG) allow for maximum detection sensitivities of 90.63%, 91.60%, and

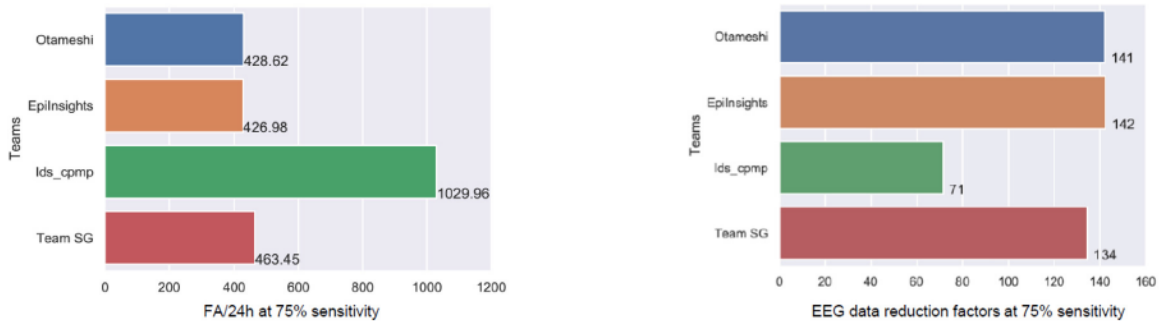

Fig. 4. An engineering step introducing a hyperparameter which allowed a trade-off between sensitivity and FA rate was included in the submissions of teams Otameshi and Ids_comp. This engineering step was applied to all 5 final submissions 4 of which thereby reached sensitivities of 75% or higher. (a) shows false alarm rates at the 75% detection sensitivity mark for those 4 models. (b) shows the reduction factors of raw EEG data that has to be reviewed by human annotators for each system. Team EpiInsights achieves the highest reduction factor of 142x.

91.57%, respectively (Fig. 7a) with data reduction factors of 28, 24 and 22 respectively. This reduces 24 h of raw data to a 51.4min-long raw data segment to be reviewed by the human annotator (Fig. 7b). Note that seizure ground truths will fluctuate across patients and over time which in turn causes fluctuating EEG data reduction factors. Hence, the developed assistive labelling systems will have the strongest annotation time saving impact for situations in which seizures are rare (short seizure ground truth) and normal brain activity is prevalent (large annotation overhead). Table 2 provides a summary of the performance parameters for the final valid submissions of all teams.

Throughout various crowdsourcing challenges, it has been observed that aggregating predictions from multiple algorithms improves over the best individual algorithm [37,38], a technique known as ensemble learning in the ML literature. The success of ensembles depends on various factors including the diversity and performance of individual algorithms [39]. We constructed several ensembles such as majority vote and the recent SUMMA algorithm [39] to evaluate all valid final submissions and compared their performance with the individual submissions. However, none of the ensembles performed better than the best individual submission in the ensemble. We mainly attribute this to the number of algorithms used for ensemble learning (5 algorithms) and the lack of sufficient diversity between these algorithms which is partly due to the fact that all the teams used the same training data.

4. Discussion

We developed and tested a novel cloud-based platform for running crowdsourced artificial intelligence challenges. The platform uses a model-to-data technique to prevent the solver community from downloading or directly accessing the challenge data while at the same time offering a notebook framework for developing models and a suite of machine learning and data pre- and post-processing tools.

Running the crowdsourced Deep Learning Epilepsy Detection Challenge in collaboration with Temple University, we enlisted a total of 87 scientists and software engineers from 14 research centres around the world to build deep learning models for automatically

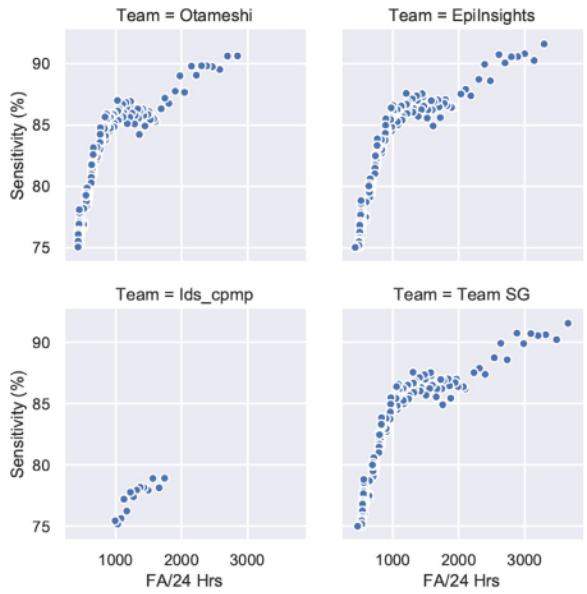

Fig. 5. FAs per 24 h plotted against detection sensitivity going from 75% sensitivity level to the maximum achievable sensitivity for each algorithm. The TAES metric causes the maximum achievable sensitivity for the model of team Ids_cpmp to stay below 80%.

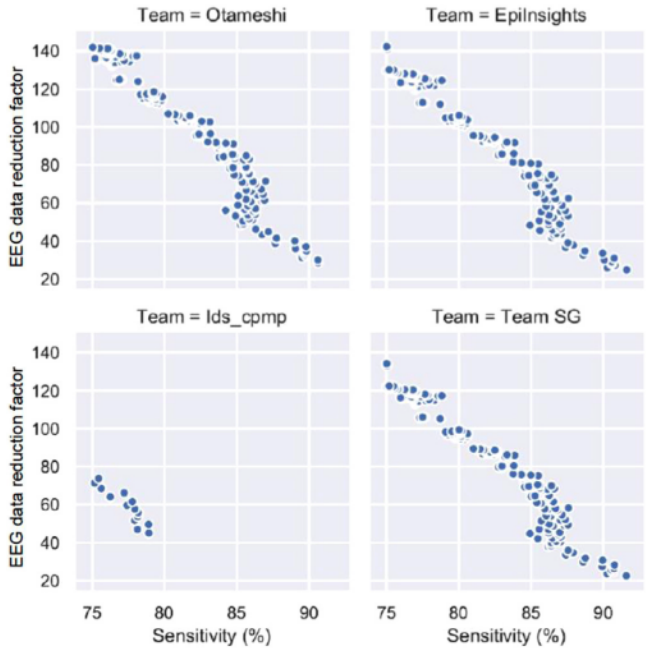

Fig. 6. Reduction factors of raw EEG data to be reviewed by a human annotator vs. detection sensitivity going from 75% to maximum achievable sensitivity values for each system. The models from teams Otameshi, EpiInsights and Team SG achieve maximum detection sensitivities of 90.63%, 91.60%, and 91.57%, respectively and two-order of magnitude data reduction factors.

Please cite this article as: S. Roy et al., Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data, EBioMedicine (2021), https://doi.org/10.1016/j.ebiom.2021.103275

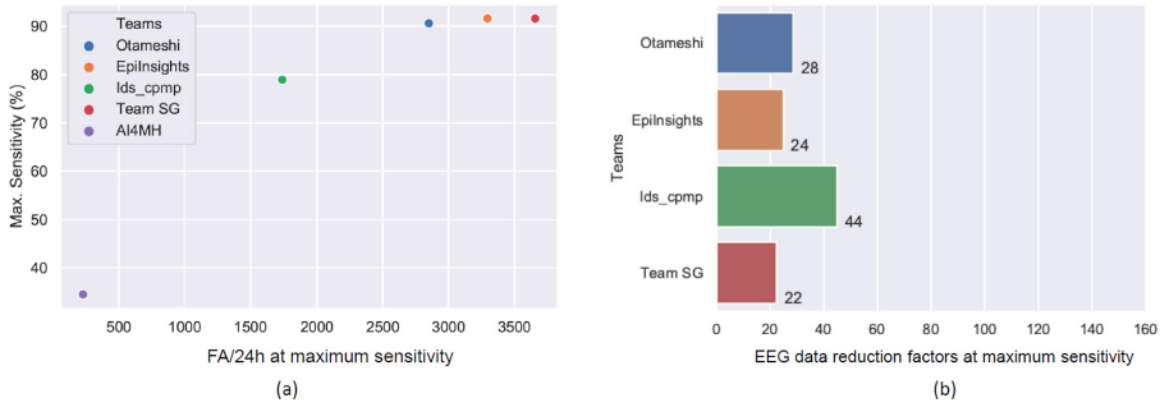

Fig. 7. (a) Applying the engineering step introduced by teams Otameshi and Ids_cpmp raises the maximum detection sensitivities to 90.63%, 91.60% and 91.57%, respectively. This comes at the cost of increased false alarm rates and decreased data reduction factors which are shown in (b). Note that even at maximum sensitivity level the lowest data reduction factor (22, Team SG) still allows to compress 24 h of raw EEG data down to a ~1h-short segment of raw EEG data to be reviewed by a human annotator.

detecting seizures in the largest existing corpus of electroencephalography (EEG) data. The best performing models demonstrated the feasibility of an assistive EEG annotation tool that could reduce the amount of raw EEG data to be reviewed by human experts by a factor of 142x thus promising to substantially decrease the time and cost burden to keep digital disease diaries.

In this section we discuss the two core aspects of this study: (i) the performance of the model-to-data crowdsourcing AI platform, and (ii) the assistive automatic EEG annotation system which was produced as part of the crowdsourced Deep Learning Epilepsy Detection Challenge.

Investments by enterprises, medical institutions and academic organizations operating in the healthcare and life sciences sector regularly result in the generation of datasets which carry substantial information content and therefore have substantial monetary and strategic value. These datasets are often large, unstructured, and noisy which makes them uniquely primed for analysis through artificial intelligence technology. However, the abundance of data is not matched by an equally strong supply of data science resources capable of developing and applying AI to drive insights from the data. Crowdsourcing the analysis of the data can solve this resourcing problem and at the same time accelerates speed, innovation and broad reproducibility of AI solutions, a benchmarking feature which the medical AI field is in dire need of [11].

As data owners intend to protect the value of their data, they are not willing to share it with open communities of solvers, ruling out the use of conventional 'Kaggle-style' AI crowdsourcing ecosystems which make challenge data directly available to the solver community. In the absence of an alternative collaborative infrastructure, many such datasets remain proprietary and unavailable for crowdsourced analysis and public benchmarking. In an effort to circumvent the need to publicly share their data and still be able to use conventional crowdsourcing platforms, some data owners have resorted to using redacted data for enabling external crowdsourced challenges [40] which generally compromises the quality of the model solutions. In other scenarios companies may use conventional crowdsourcing platforms internally [41] but in these cases, they exclusively rely on internal data scientist resources which limits size and efficiency of the solver community substantially and inhibits transparency and external verifiability of results.

Our model-to-data crowdsourcing challenge platform overcomes these limitations by allowing participants to publicly build, test, evaluate and validate AI models on proprietary data while at the same time avoiding the need to grant them access to the data itself. The novelty of our platform lies in the fact that all steps and resources required from learning about the scientific use case and challenge design to performing data pre-processing, AI model development, testing, optimisation, and submission are fully integrated in one coherent workflow, eliminating all infrastructural and procedural overhead that is not related to developing AI models. The most important platform capability is the IBM Watson Studio ecosystem which automatically provisions all compute resources through the IBM Watson Machine Learning service and all data management resources through the IBM Cloud Object Storage service. Watson Studio also leverages the Jupyter notebook framework which provides a ready-made AI coding infrastructure for data scientists. This layer of automated operational management which allows challenge participants to exclusively focus on model development and relieves them of any other operational tasks is a key advantage and novelty of our platform over conventional Kaggle-style platforms.

The platform enables collaboration between data scientists whilst keeping proprietary or sensitive data secure and protected. Our platform accomplishes this by employing a model-to-data approach in which the challenge datasets are never directly accessed by the participants who instead create models compliant with the formatting of the data based on a small sample data provided by the data

Table 2
Overview of performance parameters achieved by the final models against the blind held-out test dataset after applying the engineering step introduced by team Otameshi. The far-right column lists the minimum achievable net amount of false positive data segments (annotation overhead) which each model produces at 75% detection sensitivity and which need to be reviewed by human experts together with the correctly detected true positives (seizure ground truth) for AI-assisted manual EEG labelling.

	False Alarms/24 h at 75% Sensitivity	Raw EEG time reduction at 75% Sensitivity [factor X]	Maximum Sensitivity [%]	False Alarms/24 h at maximum Sensitivity	Raw EEG time reduction at maximum Sensitivity [factor X]	Minutes of raw EEG data to review per 24 h recording [min]
Otameshi	428.616	141.961	90.6307	2850.63	28.509	7.1436
EpiInsights	426.979	142.344	91.6025	3295.46	24.86	7.11631
Ids_cpmp	1029.96	71.4071	78.9233	1742.09	44.951	17.1661
Team SG	463.454	134.275	91.5703	3657.69	22.5135	7.72423
AI4MH	NaN	NaN	34.4671	228.027	211.751	NaN

Please cite this article as: S. Roy et al., Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data, EBioMedicine (2021), https://doi.org/10.1016/j.ebiom.2021.103275

owners. They then submit their sample models to a repository, residing within a secure cloud environment which is inaccessible to participants. There, and shielded from participants, the submitted models are trained and evaluated on the hidden data. Model performance is determined based on a pre-defined evaluation metric and the results are handed back to the respective participants. Following this scheme, the model-to-data challenge platform keeps the data shielded behind a firewall at all times while facilitating model ingestion into the model evaluator and extraction of model performances out of it. We have demonstrated and tested the first working instance of our model-to-data platform with the Deep Learning Epilepsy Detection Challenge. Further work will focus on platform upgrades through additional features for increased data safeguarding and HIPAA compliance. We plan to opensource the platform and run regular crowdsourced deep learning challenges.

The annotation models developed as part of the Deep Learning Epilepsy Detection Challenge by teams Otameshi, EpiInsights, Ids_cpmp and Team SG are capable of automatically filtering ictal segments out of raw EEG data with sensitivities that are comparable to human experts. Reaching this sensitivity regimen comes at the cost of a higher false alarm rate which, since it is substantially higher than the number of true positive samples, requires human experts to manually review all samples which the models detects for final annotation. Using these AI models as assistive filtering tools allows human data reviewers to cut down the amount of raw data that needs to be reviewed by up to two orders of magnitude. Only the collaborative combination of an automatic AI model and a human expert decision maker allows improvement of the efficiency of the EEG review and labelling process. This is a common example of how AI technology enters the realm of real-world applications: AI does not replace the human expert but rather serves as an assistive tool that enables faster and more efficient decision making.

Note that neither one of the four top performing models nor ensembling versions of the models outperform all others. For example, the model of team EpiInsights yields the highest overall achievable sensitivity of 91.60% and largest data reduction factor of 142x at 75% sensitivity but it is the model of team Ids_cpmp that produces the highest data reduction factor of 44x at maximum sensitivity. The tunability of the system is key to its deployment configuration: the choice of analytical models depends on the target sensitivity level of the overall review and the amount of time which the human reviewer is willing to invest in the final review step.

It is also important to note that we do not derive a quantitative statement on how much time exactly human reviewers will save using the developed automatic detection models. Data review processes, protocols and routines differ across institutions as do the experience and labelling performance levels of human reviewers. Furthermore, seizure frequencies per 24 h vary across patients and over the course of monitoring time windows, and the more ictal samples a 24 h raw data segment contains, the less room for raw data compression there is. These factors all affect the impact of using the automatic filtering system on the net time savings of human reviewers. Therefore, for this study we chose the net amount of raw EEG data that has to be reviewed by human annotators as a common parameter to assess the workload reduction which our system offers. Future work will test the applicability and benchmark the performance and generalisability of our automatic detection system across a variety of real-world clinical settings.

Besides integrating our models into clinical processes as assistive annotation tools of historic data, future work will also focus on further reducing false positive rates while maintaining the sensitivity levels reported in this study. If false positive rates can be reduced to human levels of 1FA/24 h [36] then the model could be used as a real-time seizure alert system.

There exists a plethora of metric frameworks for assessing the seizure detection performance of machine learning models which, although they often use similar terminology, do not allow direct performance comparison of the respective algorithms. An analysis of all popular performance metrics is beyond the scope of this paper and has been done elsewhere [24] (pre-print). However, we provide a simple example to illustrate this point: in a first scenario a deep learning model is used to detect the occurrence of a seizure event which is defined by seizure onset and end times. In a second scenario a deep learning model is used to detect seizure durations in the very same dataset. Both scenarios will describe the employed algorithms as seizure detection models and might even use the same statistical parameters to report on their performance. However, in the first scenario the algorithm will only have to detect one single ictal data sample within a seizure segment to claim success. In the second scenario the success of the algorithm will depend on how many ictal samples it detects correctly within a seizure segment. Awareness of this context and the underlying use case is crucial for being able to meaningfully compare the performance of machine learning models and to choose an appropriate validation metric for an experiment in the first place.

In this study we applied the Time-Aligned Event Scoring (TAES) metric which has been custom developed to assess the performance of detection algorithms in scenarios where both, detecting the number of events and their duration are equally important. Therefore, and in order to allow meaningful benchmarking, we stayed within the TAES framework whenever comparing the performance of models described in this study against state-of-the-art technology.

Several Kaggle or Kaggle-style AI challenges on detecting [42] and forecasting [43] epileptic seizures using EEG data have been held in the past. While these challenges also followed the crowdsourcing approach, they differ substantially from the challenge we report on in this paper with respect to management and type of challenge data as well as the obtained performances of winning models. To facilitate AI model development and data processing experiments the challenge organizers made all challenge data directly available to participants for both challenges. For the Kaggle challenge, combined intracranial EEG datasets from humans and dogs were used as challenge data whereas we used exclusively non-invasive human scalp EEG data in our challenge. The Kaggle-style Neureka challenge employed a scalp EEG dataset and the TAES scoring metric but tasked participants to develop AI models for forecasting seizures (i.e. unlike detection, predicting them before they occur) and to at the same time minimize the number of EEG channels. The winning model showed a human-level FA rate of 1.44/24 h but also a sensitivity of 12.37% which prevents the model from being suitable for real-life clinical applications [44].

Future work will focus on assessing the suitability of our assistive EEG annotation system in real-world clinical settings and on upgrading and open sourcing our model-to-data crowdsourcing AI challenge platform based on the insights we gained from running the Deep Learning Epilepsy Detection Challenge.

5. Contributors

Study design and challenge organization: SR, IK, MM, TM, AB, JT, TS, JTa, UA, JP, IO, BDAM, SM, RK, MRZ, GS, SH.

Challenge platform design and implementation: SR, IK, MM, TM, AB, JT, TS, BDAM, SM, RK, MRZ, GS, SH.

Data collection: JP, IO.

Data preparation: SR, IK, TS, JP, IO, SM, MRZ, GS, SH.

Development of analytical AI models and data pre-and post-processing models: The IBM Epilepsy Consortium: TI, HY, HP, PM, YQ, SHA, RC, WL, WLT, LM, MC, DAD, PH.

Model evaluation and interpretation of results: SR, MEA, GS, SH.

Paper writing: SR, IK, MM, MEA, JP, MRZ, GS, SH.

Generating figures: SR, MM, JTa, SH.

Please cite this article as: S. Roy et al., Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data, EBioMedicine (2021), https://doi.org/10.1016/j.ebiom.2021.103275

All authors read and approved the final version of the manuscript. The corresponding authors Stefan Harrer and Gustavo Stolovitzky declare that they had full access to all the data in the study and that they had final responsibility for the decision to submit for publication.

Declaration of Competing Interest

SR, IKK and SH are inventors on issued patent US 10,596,377. HY is an inventor on pending patent US 16/670,177. All other authors do report no conflicts of interest.

Acknowledgments

We would like to thank Carlos Fonseca and the IBM Cloud Team for support with setting up cloud accounts for challenge participants as well as Elise Blaese for guidance in designing the challenge launch plan, Olivia Smith for mathematical guidance, Josh Andres for help with designing the web portal and John Cohn for sharing his experience with client data management systems. IBM employed all IBM Research authors. Temple University employed all Temple University authors. The Icahn School of Medicine at Mount Sinai employed Eren Ahsen.

Data Sharing: All data that underlie the results reported in this article, after anonymization (text, tables, figures, supplemental information) is available publicly as open source data at the website https://www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml. All data underlying the design of the developed analytical models is available publicly through the supplemental information of this article. All data is available immediately with publication. We plan to open source the challenge platform after completion of a public challenge which is ongoing at the time of publication (https://www.ibm.com/blogs/research/2020/12/object-recognition-models/).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2021.103275.

References

[1] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19(6):1236–46.
[2] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8.
[3] Zhu W, Xie L, Han J, Guo X. The application of deep learning in cancer prognosis prediction. Cancers 2020;12(3):603. (Basel)Mar 5.
[4] Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572:116–9.
[5] 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.
[6] Harutyunyan H, Khachatrian H, Kale DC, et al. Multitask learning and benchmarking with clinical time series data. Sci Data 2019;6:96.
[7] Nurse E, Mashford BS, Yepes AJ, Kiral-Kornek I, Harrer S, Freestone DR. Decoding EEG and LFP signals using deep learning: heading TrueNorth. In: Proceedings of the ACM International Conference on Computing Frontiers; 2016. p. 259–66.
[8] Kiral-Kornek I, Roy S, et al. Epileptic seizure prediction using big data and deep learning: toward a mobile system. EBioMedicine 2018;27:103–11.
[9] Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. Nat Biotechnol 2018;36:391–2.
[10] Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3(3):e200265.
[11] Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2020;323(4):305–6.
[12] Epilepsy Foundation. About Epilepsy: The Basics. Bowie, Maryland, USA: Epilepsy Foundation; 2021. [cited 2021 Jan 30]. Available from: https://www.epilepsy.com/learn/about-epilepsy-basics.
[13] Saab K, Dunnmon J, Ré C, et al. Weak supervision as an efficient approach for automated seizure detection in electroencephalography. NPJ Digit Med 2020;3:59.
[14] Mirmomeni M, Fazio T, von Cavallar S, Harrer S, et al. Wearable Sensors – Fundamentals, Implementation And Applications. 2nd Edition Academic Press; 2020 ISBN 9780128192467.
[15] Harrer S, et al. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019;40(8):577–91.
[16] Fisher RS, Blum DE, DiVentura B, et al. Seizure diaries for clinical research and practice: limitations and future prospects. Epilepsy Behavior 2012;24(3):304–10.
[17] Siddiqui MK, et al. A review of epileptic seizure detection using machine learning classifiers. Brain Inform 2020;7(1):5. Published online 2020 May 25.
[18] BioSpace. Epilepsy Monitoring Devices Market: Advances In Wearable Technology Show Promise In Preventing Epileptic Seizures. Albany, NY, USA: BioSpace; 2020. [updated 2020 Sept 9; cited 2021 Jan 30]. Available from: https://www.biospace.com/article/epilepsy-monitoring-devices-market-advances-in-wearable-technology-show-promise-in-preventing-epileptic-seizures/.
[19] Ofri D. Perchance to Think. N Engl J Med 2019;380:1197–9.
[20] Juhász C, Berg M. Computerized seizure detection on ambulatory EEG: finding the needles in the haystack. Neurology 2019;92(14):641–2.
[21] González Otárula KA, Mikhaeil-Demo Y, Bachman EM, Balaguera P, Schuele S. Automated seizure detection accuracy for ambulatory EEG recordings. Neurology 2019;92(14):e1540–6.
[22] Shah V, von Weltin E, Lopez S, McHugh JR, Veloso L, Golmohammadi M, Obeid I, Picone J. The Temple university hospital seizure detection corpus. Front Neuroinform 2018;12:83.
[23] Scheuer ML, Wilson SB, Antony A, Ghearing G, Urban A, Bagić AI. Seizure detection: interreader agreement and detection algorithms assessments using a large dataset. J Clin Neurophysiol 2020. doi: 10.1097/wnp.0000000000000709.
[24] Ziyabari S., Shah V.L., Golmohammadi M., Obeid I., Picone J. Objective evaluation metrics for automatic classification of EEG events. 2019; arXiv:1712.10107 [cs.LG].
[25] Temple University. Open Source Eeg Resources. Philadelphia, PA, USA: The Neural Engineering Data Consortium; 2021. [updated 2021 Jan 7; cited 2021 Jan 30]. Available from: https://www.isip.piconepress.com/projects/tuh_eeg/.
[26] Obeid I, Picone J. The Temple university hospital EEG data corpus. Front Neurosci 2016;10:196.
[27] IBM Cloud. Watson studio. New York, NY, USA: IBM Cloud; 2020. [updated 2020 Dec 6; cited 2021 Jan 30]. Available from: https://cloud.ibm.com/catalog/services/watson-studio.
[28] Perkel JM. Why Jupyter is data scientists' computational notebook of choice. Nature 2018;563(7732):145–7.
[29] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th Symposium on Operating Systems Design and Implementation; 2016. p. 265–83.
[30] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017.
[31] IBM Cloud. IBM Cloud Object Storage. New York, NY, USA: IBM Cloud; 2021. [cited 2021 Jan 30]. Available from: https://www.ibm.com/cloud/object-storage.
[32] Maetschke S., Tennakoon R., Vecchiola C., Garnavi R. Nuts-flow/ml: data pre-processing for deep learning. arXiv preprint arXiv:1708.06046. 2017.
[33] Homan RW. The 10-20 electrode system and cerebral location. Am J EEG Technol 1988;28(4):269–79.
[34] Jayant NA, Abeer JH, Janna C, Parthasarathy T, Tammy NT. American clinical neurophysiology society guideline 2: guidelines for standard electrode position nomenclature. Neurodiagn J 2016;56(4):245–52.
[35] López de Diego S. Automated interpretation of abnormal adult electroencephalograms [temple university]. 2017. ProQuest Dissertations and Theses. Available from: https://search-proquest-com.libproxy.temple.edu/pqdtlocal1005760/docview/1950580989/7C485ED2A6F443A9PQ/1?accountid=14270.
[36] Golmohammadi M, Shah V, Obeid I, Picone J. Deep learning approaches for automatic seizure detection from scalp electroencephalograms. In: Obeid I, Selesnick I, Picone J, editors. Signal Processing In Medicine And Biology: Emerging Trends In Research And Applications. 1st edition Springer; 2020. p. 233–74.
[37] Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, Yu T, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat Commun 2019;10(1):1–17.
[38] Choobdar S, Ahsen M, Crawford J, Tomasoni M, Lamparter D, Lin J, Hescott B, et al. Open community challenge reveals molecular network modules with key roles in diseases. Nat Methods 2018. doi: 10.1101/265553.
[39] Ahsen ME, Vogel RM, Stolovitzky GA. Unsupervised evaluation and weighted aggregation of ranked classification predictions. J Mach Learn Res 2019;20(166):1–40.
[40] AustralianMining. OZ minerals, unearthed award $1 m prize for exploration contest. Melbourne, VIC, AU: Australian Mining; 2019 [updated 2019 July 1, cited 2021 Jan 30]. Available from: https://www.australianmining.com.au/news/oz-minerals-unearthed-award-1m-prize-for-exploration-contest/.

[41] Forbes. Why the new open data initiative by Microsoft. Adobe and sap could revolutionize customer experience. Jersey City, NJ, USA: Forbes; 2018. [updated 2018 Sept 28, cited 2021 Jan 30]. Available from: https://www.forbes.com/sites/ryanholmes/2018/09/28/why-the-new-open-data-initiative-by-microsoft-adobe-and-sap-could-revolutionize-customer-experience/#14b4095952e4.
[42] Baldassano SN, Brinkmann BH, Ung H, et al. Crowdsourcing seizure detection: algorithm development and validation on human implanted device recordings. Brain 2017;140(6):1680–91.
[43] NeuroTechX. Neureka Epilepsy Challenge. Philadelphia, PA, USA: NeuroTechX and Novela Neurotech; 2020. [cited 2021 Jan 30]. Available from: https://neureka-challenge.com/#data.
[44] Chatzichristos C, Dan J, Narayanan M, Seeuws N, Vandecasteele K, De Vos M, Bertrand A, Van Huffel S. Epileptic seizure detection in EEG via fusion of multi-view attention-gated U-net deep neural networks. In: Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB) 2020; 2020.

Please cite this article as: S. Roy et al., Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data, EBioMedicine (2021), https://doi.org/10.1016/j.ebiom.2021.103275