Evaluation of Combined Artificial Intelligence and Neurologist Assessment to Annotate Scalp Electroencephalography Data

Subhrajit Roy[1,*,o], Isabell Kiral-Kornek[1,€,o], Mahtab Mirmomeni[1,o], Todd Mummert[2], Alan Braz[2], Jason Tsai[2], Jianbin Tang[1], Umar Asif[1], Thomas Schaffter[8], Mehmet Eren Ahsen[10,†], Toshiya Iwamori[4], Hiroki Yanagisawa[4], Hasan Poonawala[5, ¥], Piyush Madan[6], Yong Qin[7], Joseph Picone[9], Iyad Obeid[9], Bruno De Assis Marques[1], Stefan Maetschke[1], IBM Epilepsy Consortium[11], Rania Khalaf[6], Michal Rosen-Zvi[3], Gustavo Stolovitzky[2,©], Stefan Harrer[1,©]

[1]IBM Research – Melbourne, Australia
[2]IBM Research – T. J. Watson Research Center, USA
[3]IBM Research – Haifa, Israel.
[4]IBM Research – Tokyo, Japan
[5]IBM Research – Kuala Lumpur, Singapore
[6]IBM-MIT AI Lab, Cambridge, USA
[7]IBM Research – Beijing, China
[8]Sage Bionetworks – Seattle, USA
[9]Temple University Hospital - Philadelphia, USA
[10]Icahn School of Medicine at Mount Sinai, NY, USA

[*] Google Brain, London, UK
[€] Amazon Web Services, Melbourne, AU
[¥] Amazon Web Services, London, UK
[†]Department of Business Administration, University of Illinois at Urbana-Champaign, Champaign, IL, USA

[o] *These authors contributed equally.*

[©] *Corresponding authors:* gustavo@us.ibm.com, sharrer@au.ibm.com

[11]The IBM Epilepsy Consortium has the following members: Toshiya Iwamori, Hiroki Yanagisawa, Hasan Poonawala, Piyush Madan, Yong Qin, Sharon Hensley Alford, Rachita Chandra, Wen Liu, Wei Lian Ti, Li Ma, Michael Cherner, Dario Arcos-Diaz, Paul Hake

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 minimising 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 171,745s ictal and 2,185,864s 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 to reduce 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 optimisation 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 Hospital authors. The Icahn School of Medicine at Mount Sinai employed Eren Ahsen. 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.

Keywords

Epilepsy, Seizure Detection, Artificial Intelligence, Deep Neural Networks, EEG, Automatic Labelling, Crowdsourcing Challenges

Research in Context

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 which can empower doctors to optimize and adjust medication response and pharma to design more efficient clinical trials. Until recently such disease diaries were kept fully manually making data review a highly cost- and time-intense process when performed by medical experts in clinical settings and rendering records highly inaccurate when populated by patients through self-reporting outside the clinic. A new way to overcome the challenges of manual data review was opened by recent breakthroughs in deep learning techniques allowing to build models for automatic detection of seizures in brain activity data monitored by electroencephalography (EEG) sensors. Harnessing the power of crowdsourced artificial intelligence algorithm development and using one of the largest EEG datasets in existence we have demonstrated an automatic seizure detection model which can assist human reviewers in substantially cutting down the amount of raw EEG data to be annotated manually. 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% and 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 reaching from cancer diagnosis [2][3], prediction of acute kidney injury [4], detection of diabetic retinopathy [5], mining of electronic health records [6] to the development of brain-machine-interfaces [7][8]. While Kaggle has pioneered the crowd-sourcing of machine learning challenges to incentivize data scientists from around the world to advance algorithm and model design, the increasing complexity of problem statements 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 match this description are few and far between, unfortunately leading to a shrinking 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 allowing to run a crowdsourced challenge with it, we have recently introduced a challenge ecosystem that utilizes the so-called model-to-data paradigm [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 before providing back model performance results to participants. Solvers can then use the model performance to improve their algorithms. In this scheme participants, although not being given the option to download or directly access the challenge data at any point have the full suite of crowdsourced challenge tools at their disposal. This challenge platform 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 platform, recently, we 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 individualised condition, its expression changes from patient to patient, and even for a specific patient pathological patterns 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 of 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 so called 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 optimisation 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. Entirely depending on third party or self-reporting, manual disease diaries are only approximately 50% accurate [16]. This is not rooted in sloppy reporting techniques: it is the individualised and incapacitating nature of the disease itself that leaves patients unable to recognise, remember or keep track of their own seizures and that makes it impossible for medically untrained observers to recognise and describe seizure episodes in clinically actionable ways [15]. In order to overcome this challenge and leveraging the advent of a plethora of wearable and mobile sensing platforms the field has turned to exploring the use of machine learning techniques for developing automatic patient monitoring and seizure detection systems [17]. Amongst a broad spectrum of sensor modalities ranging from video

cameras to smart watches [18], electrodes measuring brain activity in form of electroencephalography (EEG) data are considered to be the gold standard for seizure monitoring [13]. However, while EEG monitoring systems have evolved from relying on intracranial implanted electrodes to making use of non-invasive wearable devices the automatic annotation of EEG data remains a challenging machine learning problem. Amongst the main reasons for this are device related issues such as a low signal to noise ratio, movement artefacts, poor electrical conduction, and nonlinearly distorted crosstalk between spatially adjacent sensors as well as disease specific intricacies such as the highly individualised profiles of seizure patterns which make generalisability of detection models across patients a difficult endeavour. As a result, in today's clinical practice review and annotation of EEG data is done manually by trained neurologists. The time and cost burdens of this routine are substantial: they account for approximately 5% of total hospital charges for epilepsy patients admitted to 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] or of having to reduce the duration of monitoring sessions to cut down the amount of EEG data to be reviewed [13].

Aiming to relief medical experts from the need of manual EEG review, a variety of machine learning-based automatic EEG annotation systems have been proposed [13][20], and some of them have been deployed and tested in clinical scenarios [21][22]. While limited practical value of some commercially available automatic EEG annotation tools could be demonstrated [23], the lack of commonly adopted performance metrics to evaluate them against human expert reviewers [24] and of generalisability across datasets collected using different measuring setups and institutions [13] has inhibited broad adoption of automatic annotation systems in critical care settings.

Using the world's largest EEG dataset, the TUH Seizure Corpus [25], 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 the developed annotation models for being used 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 the goal to run a challenge that mobilizes the largest possible pool of participants globally across IBM, we designed a crowdsourced challenge, the Deep Learning Epilepsy Detection Challenge, in which 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 generously provided by Temple University Hospital (TUH) [22][26]. We partition this data to create a training, validation and test sets which participants could access only through our platform.

In order 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 under 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.

In addition, our platform further 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.

Study Design

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 Figure 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, and scientific background. 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. **(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 necessary python libraries and custom functions for the invisible integration with **(4) IBM's Cloud Object Storage (COS)** [31] and the analytics engine **(5) Watson Machine Learning (WML)**. In dedicated spots in the notebook, participants could write custom pre-processing code, 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 (GPUs). Code was bundled up automatically in the starter kit and deployed to and run 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. COS held the data for this challenge. Note that using the starter kit, participants submitted their model code to the platform which autonomously organized model training and testing on the raw data and provided back model performance results to participants who could then investigate this feedback in order to better design custom algorithms. This approach is called the 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-process data into a more common format, to optimize streaming through the use of the NutsFlow and NutsML libraries [32], and to provide seamless access to all services used. **Final code scoring** after completion of the challenge was conducted in an automated way as soon as code was submitted through the starter kit.

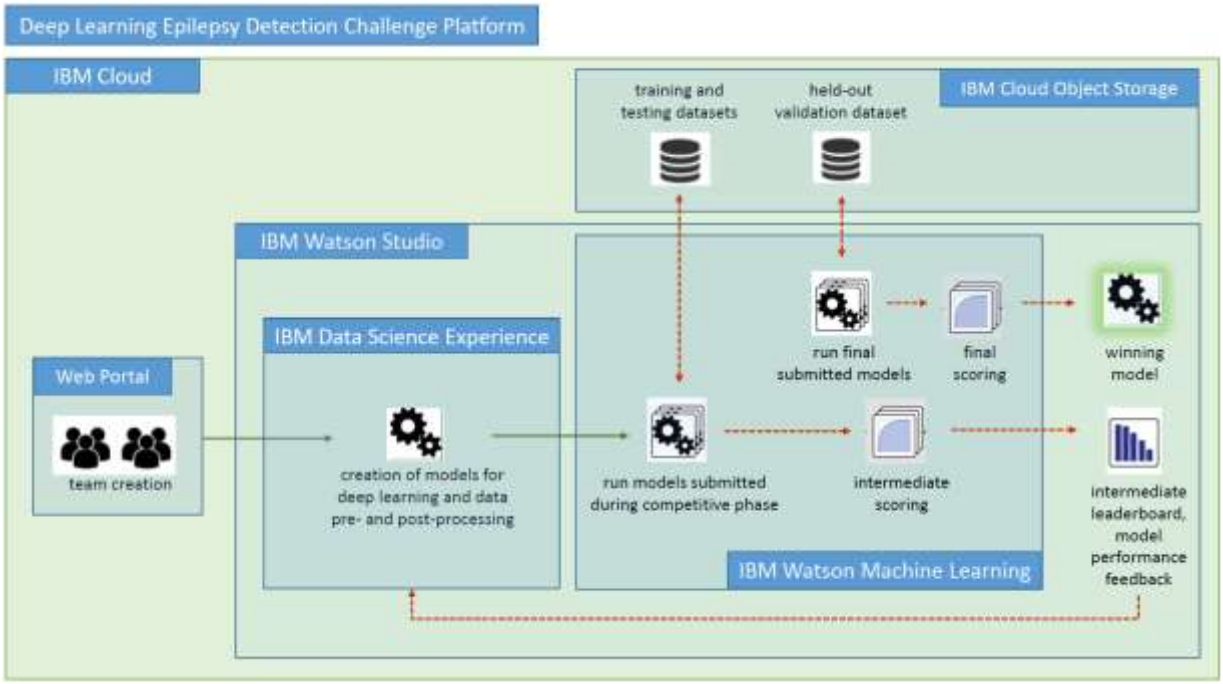

Figure 1: *High-level architecture of the custom-built challenge platform depicting data and model flow during challenge operation following the model-to-data paradigm: challenge participants do at no point download or access the data directly. Instead they create and submit models to the platform (green solid arrows) which automatically organises training and testing behind a secure firewall and then provides back model performance results to participants (orange dashed arrows). This is fundamentally different to conventional crowdsourced challenge setups.*

Data sources and preparation

The TUH EEG Seizure Corpus v1.2.0 [22] which contains scalp EEG records of 315 patients with annotated seizure times was split and used for training and validation datasets for the challenge (Table 1). The validation dataset was used to determine team rankings on the intermediate leaderboard during the competitive phase (Figure S1). 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) and was used for final team rankings on the final leaderboard at the end of the challenge (Figure S1). 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 to train and evaluate the participants' models following the model-to-data paradigm described above. The blind, held-out test set was not accessible to participants' models at any time during the challenge and was only used only once by the challenge organising team to evaluate the submitted models during the scoring phase after the completion of the competitive phase (supplemental information).

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 [Acharya2016]. 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].

	Training set	Validation set	Blind test set
Patients	265	50	50
EDF files	2032	1032	1022
Seizure (secs)	76517	55764	39464
Non-seizure (secs)	1119863	562331	503670
Total (secs)	1196381	618096	543134

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

Evaluation procedure

The evaluation of machine learning algorithms for seizure detection lacks standardization, since there is no agreed-upon standard metric for evaluation in this specific community. 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 Hours, 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) first we gave participants a 2-week time window to submit their final trained model. Next, from each final submission we extracted the pre-processing model and post-processing code and ran them on a held-out blind test dataset (which participants never 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.

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, hence it was these 5 teams named Ids_cpmp, Otameshi, AI4MH, Team SG, and EpiInsights that entered the final evaluation stage. Evaluation metric E, Sensitivity S, False alarm rate FA/24h obtained on the validation (leaderboard) and the blind test set and a Sensitivity vs. FA/24h plot for all 5 submissions are provided in Figure 2. It can be seen that the performance of the 5 submissions is similar in both validation and tests set, indicating that there is no evidence of overfitting.

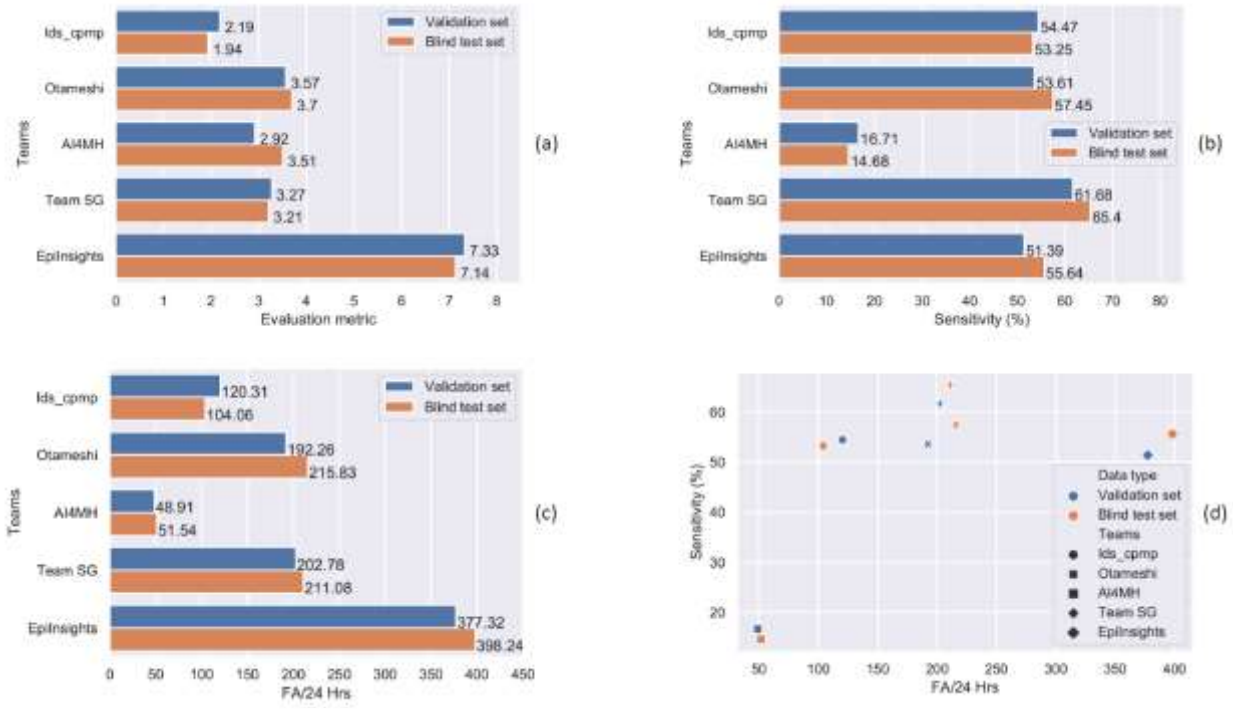

Figure 2: *All 5 valid final submissions were tested against validation and blind test sets. The plots show the results for (a) evaluation metric E, (b) sensitivity S, (c) false alarm rate FA per 24h and (c) FA/24h plotted against S.*

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 mark stands at 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, using an assisting labelling system they will only need to review and verify all raw 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). However, as part of their final solution, team Otameshi and Ids_cpmp introduced an engineering step which added synthetic false alarms (details provided in supplemental information). This step included a hyperparameter which allowed for the tuning of sensitivity and false alarm rate. We removed this engineering step for producing the results shown in Figure 2 to be able to assess detection performances 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 to increase their sensitivities to and above the 75% level threshold for all submissions except for the one from team AI4MH which is therefore excluded from the following analyses (details provided in supplemental information). The total false alarm numbers per 24h obtained by each of these four submissions at 75% sensitivity are shown in Figure 4a and yield the shortest achievable annotation overheads for each automatic seizure detection system as depicted in Figure 3.

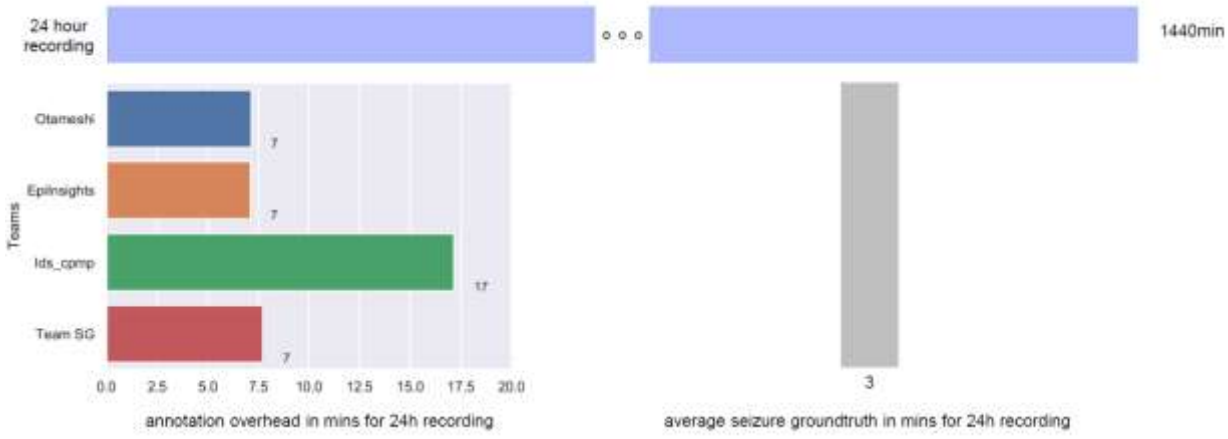

Figure 3: *In order to label 24h of EEG recordings an unassisted human annotator has to review all 24h of raw EEG data (top). Using the systems developed by teams Otameshi, EpiInsights, Ids_cpmp and Team SG a clinician will only have to review all segments which the system automatically detects, i.e. 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 24h raw EEG data*

recording is ~0.2% [22] or ~3min. The best models achieve a minimum annotation overhead of 7min which therefore allows to reduce the total amount of raw EEG data to be reviewed by a human annotator from 24h down to 10min 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.

Teams Otameshi, EpiInsights and Team SG all achieve minimum annotation overheads of 7min. In good approximation it can be assumed that on average ~0.2% or ~3min of a continuous 24h-long raw EEG recording describe ictal segments while 98.8% or 1437min of the raw data are correlated with non-ictal episodes [22]. For unassisted human labelling of 24h of raw EEG data this means that the seizure ground truth is 3min and the annotation overhead is 1437min. Using the automatic labelling systems reduces the annotation overhead to 7min thus reducing the amount of total raw EEG data that needs to be reviewed by a human expert from 24h to 10min. 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 (Figure 4b) and thus lead to a substantial decrease of the time and cost burden for all human annotation scenarios.

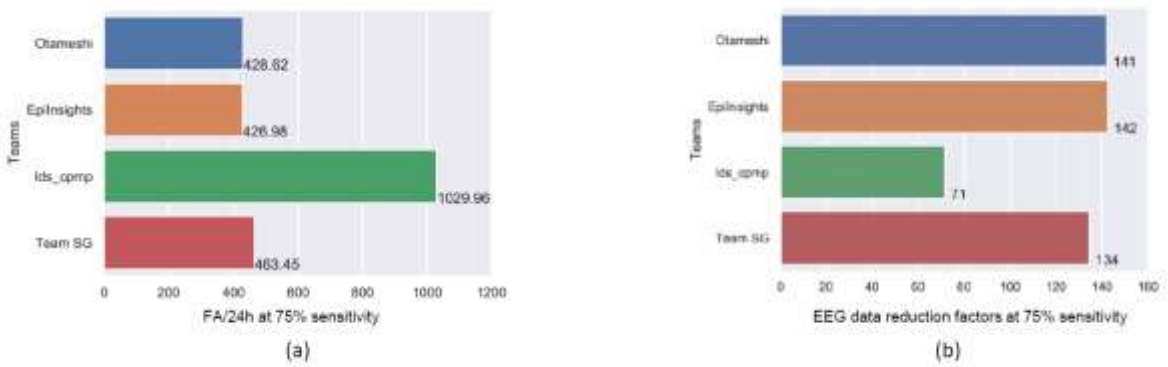

Figure 4: *An engineering step introducing a hyperparameter which allowed to tune sensitivity vs. false alarm rate was included in the submissions of teams Otameshi and IDS_comp. This engineering step was applied to all 5 final submissions 4 of which thereby reached sensitivities of 75% or higher. (a) shows false alarm rates at the 75% detection sensitivity mark for those 4 models which can be used as assistive labelling tools for raw EEG data annotation: instead of having to review the entire raw data, human annotators only have to review the events detected by the system. (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.*

Further investigating the effect of the engineering step, 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. Figure 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 (Figure 6). Exploiting this effect allows to develop 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 the three systems of teams Otameshi, EpiInsights and Team SG each allow for maximum detection sensitivities of 90.63%, 91.60%, and 91.57% respectively (Figure 7a) with data reduction factors of 28, 24 and 22 respectively allowing to reduce 24h of raw data to a 51.4min-long raw data segment to be reviewed by the human annotator (Figure 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.

	False Alarms/24h @ 75% Sensitivity	Raw EEG time reduction @ 75% Sensitivity (X)	Max. Sensitivity (%)	False Alarms/24h @ Max. Sensitivity	Raw EEG time reduction @ Max. Sensitivity (X)	Minutes of raw EEG to review for 24h recording
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

Table 2: *Overview of performance parameters achieved by the final valid submitted models of all teams against the blind held-out test dataset and 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.*

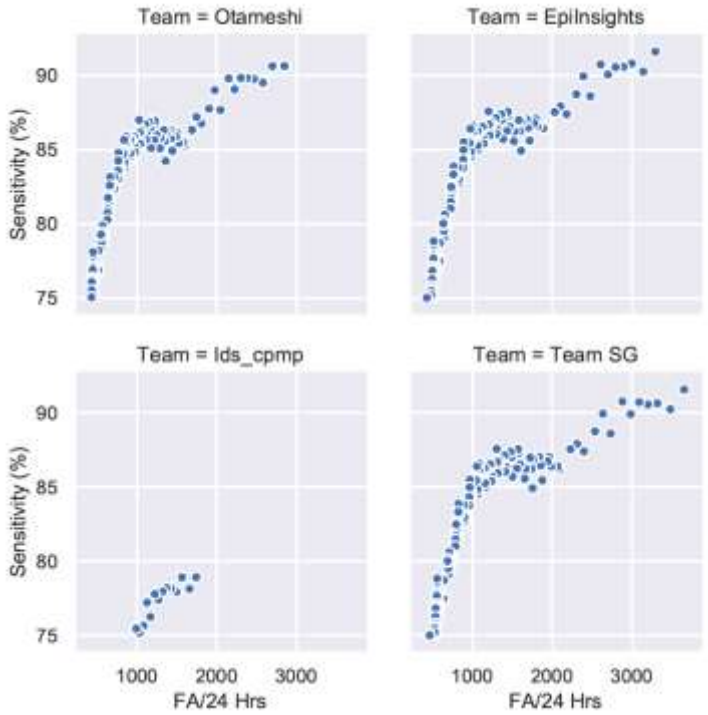

Figure 5: *False alarm rates per 24h plotted against detection sensitivity going from 75% sensitivity level to the maximum achievable sensitivity for each algorithm. Note that team AI4MH is not included since their solution could not reach the 75% sensitivity level by applying the engineering step.*

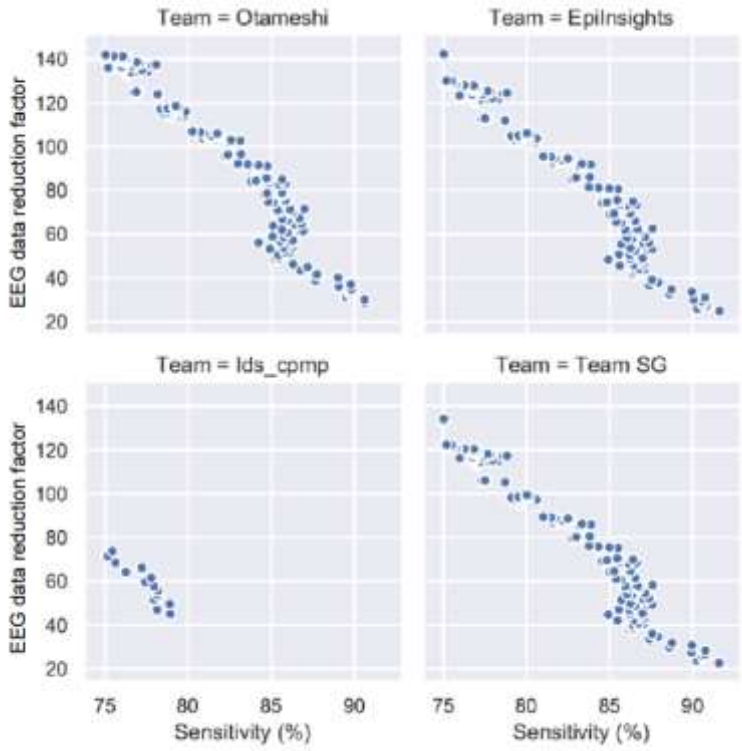

Figure 6: *Reduction factors of raw EEG data to be reviewed by a human annotator vs. detection sensitivity going from 75% to maximum achievable sensitivity values for each system. The models from teams Otameshi, EpiInsights and Team SG achieve maximum detection sensitivities of 90.63%, 91.60%, and 91.57%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

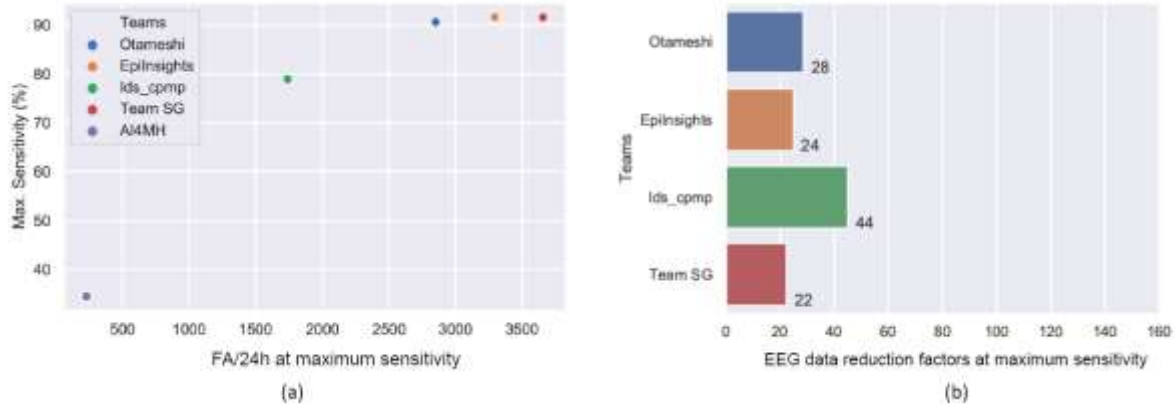

Figure 7: *(a) Applying the engineering step introduced by teams Otameshi and Ids_cpmp allows to bring the models of teams Otameshi, EpiInsights and Team SG up to maximum detection sensitivities of 90.63%, 91.60% and 91.57% respectively. This comes at the cost of increased false alarm rates and decreased 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 24h of raw EEG data down to a ~1h-short segment of raw EEG data to be reviewed by the human annotator.*

Throughout various crowdsourcing challenges, it has been observed that aggregating prediction of a set of multiple algorithms improves over the best individual [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

In this section we discuss the two core aspects of this study: (i) the performance of our custom-developed 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 organisations 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 crowd-sourcing 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 crowd-sourcing platforms, some data owners have resorted to using redacted data for enabling external crowd-sourced challenges [40] which generally compromises the quality of the model solutions. In other scenarios companies may use conventional crowd-sourcing 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 to publicly build, test, evaluate and validate AI models on proprietary data while at the same time avoiding the need to grant the solver community 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 such platform design features are 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 as well as the Jupyter notebook framework on the IBM Data Science Experience platform which provides a ready-made AI coding infrastructure for data scientists. The platform can be used to collaboratively employ data scientists from custom-defined solver communities to work together whilst keeping the proprietary or sensitive data secure and protected. Our platform accomplishes this by employing a model-to-data approach in which

the challenge data is 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 owners. They then submit their sample models to a repository of models, residing within a secure cloud environment which is inaccessible to participants. There, and shielded from participants, their submitted models are trained and evaluated on the actual data. Model performances are determined based on a pre-defined evaluation metric and the results of such evaluation runs 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 firewalled 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 by means of running the Deep Learning Epilepsy Detection Challenge. Further work will focus on platform upgrades through additional features for increased data safeguarding and HIPPA 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 non-inferior to human experts. Reaching this sensitivity regimen comes at the cost of a 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 to improve the efficiency of the EEG review and labelling process. This is indicative for 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 human 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 developed 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 24h vary across patients and over the course of monitoring time windows, and the more ictal samples a 24h 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 brought down to human levels of below 7FA/24h [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. While explaining all existing performance metric flavours would go beyond the scope of this paper and has been done

elsewhere [24] we give a simple example to illustrate the point: in a first scenario a deep learning model is used to detect the seizure frequency in a given time series data strain that contains seizure onset and end labels. 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.

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.

Highlights:

- 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 postprocessing tools.
- Running the crowdsourced Deep Learning Epilepsy Detection Challenge in collaboration with Temple University Hospital, we enlisted a total of 87 scientists and software engineers from 14 research centres around the world to build deep learning models for automatically detecting seizures in the largest existing corpus of electroencephalography (EEG) data. 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 144x.

Acknowledgements: 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 and Josh Andres for help with designing the web portal.

Funding sources: *(to be confirmed by all co-authors on authorship declaration)*

Conflicts of interest: *(to be confirmed by all co-authors on authorship declaration)*

Author contributions: *(to be confirmed by all co-authors on authorship declaration)*

References:

[1] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018 Nov 27;19(6):1236-1246. doi: 10.1093/bib/bbx044. PMID: 28481991; PMCID: PMC6455466

[2] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), p.115.

[3] Zhu W, Xie L, Han J, Guo X. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers (Basel)*. 2020;12(3):603. Published 2020 Mar 5. doi:10.3390/cancers12030603

[4] Tomašev, N., Glorot, X., Rae, J.W. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572,** 116–119 (2019).

[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):2402–2410.

[6] Harutyunyan, H., Khachatrian, H., Kale, D.C. *et al.* Multitask learning and benchmarking with clinical time series data. *Sci Data* **6,** 96 (2019).

[7] Nurse, E., Mashford, B.S., Yepes, A.J., Kiral-Kornek, I., Harrer, S. and Freestone, D.R., 2016, May. Decoding EEG and LFP signals using deep learning: heading TrueNorth. In *Proceedings of the ACM International Conference on Computing Frontiers* (pp. 259-266). ACM.

[8] Kiral-Kornek, Isabell, Roy, Subhrajit et al. "Epileptic seizure prediction using big data and deep learning: toward a mobile system." *EBioMedicine* 27 (2018): 103-111.

[9] Guinney, J., Saez-Rodriguez, J. Alternative models for sharing confidential biomedical data. *Nat Biotechnol* **36,** 391–392 (2018).

[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):305–306.

[12] 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.* **3,** 59 (2020).

[14] M. Mirmomeni, T. Fazio, S. v. Cavallar, S. Harrer et al. "*Wearable Sensors – Fundamentals, Implementation and Applications – 2nd Edition*", ISBN 9780128192467, Academic Press, in press, Nov 2020.

[15] S. Harrer et al. "*Artificial intelligence for clinical trial design*", Trends in Pharmacological Sciences (Cell Press) 2019.

[16] Fisher R S, Blum D E, DiVentura B, et al. Seizure diaries for clinical research and practice: limitations and future prospects[J]. Epilepsy & Behavior, 2012, 24(3): 304-310.

[17] Brain Inform. 2020 Dec; 7(1): 5. Published online 2020 May 25. doi: 10.1186/s40708-020-00105-1. A review of epileptic seizure detection using machine learning classifiers

[18] https://www.biospace.com/article/epilepsy-monitoring-devices-market-advances-in-wearable-technology-show-promise-in-preventing-epileptic-seizures/

[19] Danielle Ofri “Perchance to Think”, N Engl J Med 2019; 380:1197-1199 DOI: 10.1056/NEJMp1814019

[20] Computerized seizure detection on ambulatory EEG Finding the needles in the haystack Csaba Juhász, Michel Berg Neurology Apr 2019, 92 (14) 641-642;

[21] Automated seizure detection accuracy for ambulatory EEG recordings Karina A. González Otárula, Yara Mikhaeil-Demo, Elizabeth M. Bachman, Pedro Balaguera, Stephan Schuele Neurology Apr 2019, 92 (14)

[22] Shah, V., v. Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., Obeid, I. and Picone, J. 2018. The Temple University Hospital Seizure Detection Seizure Detection Corpus. *Frontiers in Neuroinformatics, 12*, p. 83.

[23] Scheuer, Mark L.[*]; Wilson, Scott B.[*]; Antony, Arun[†]; Ghearing, Gena[‡]; Urban, Alexandra[†]; Bagić, Anto I.[†] Seizure Detection, Journal of Clinical Neurophysiology: May 27, 2020

[24] Ziyabari, S., Shah, V., L., Golmohammadi, M., Obeid, I. and Picone, J. 2019. *Objective evaluation metrics for automatic classification of EEG events*. arXiv:1712.10107 [cs.LG].

[25] https://www.isip.piconepress.com/projects/tuh_eeg/

[26] Obeid, I. and Picone, J., 2016. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, *10*, p.196.

[27] https://cloud.ibm.com/catalog/services/watson-studio

[28] Perkel, Jeffrey M. "Why Jupyter is data scientists' computational notebook of choice." *Nature* 563.7732 (2018): 145-147.

[29] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (pp. 265-283).

[30] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A., 2017. Automatic differentiation in pytorch.

[31] https://www.ibm.com/cloud/object-storage

[32] Maetschke, S., Tennakoon, R., Vecchiola, C. and Garnavi, R., 2017. nuts-flow/ml: data pre-processing for deep learning. *arXiv preprint arXiv:1708.06046*.

[33] Richard W. Homan (1988) The 10-20 Electrode System and Cerebral Location, American Journal of EEG Technology, 28:4, 269-279.

[34] Jayant N. Acharya, Abeer J. Hani, Janna Cheek, Parthasarathy Thirumala & Tammy N. Tsuchida (2016) American Clinical Neurophysiology Society Guideline 2: Guidelines for Standard Electrode Position Nomenclature, The Neurodiagnostic Journal, 56:4, 245-252.

[35] Silvia López. Automated Interpretation of Abnormal Adult Electroencephalograms.MS Thesis, Tem-ple University. Link:https: // www. isip. piconepress. com/ publications/ ms_ theses/ 2017/abnormal/, 2017.

[36] Golmohammadi, M., Ziyabari, S., Shah, V., de Diego, S. L., Obeid, I., & Picone, J. (2017). Deep architectures for automated seizure detection in scalp eegs. *arXiv preprint arXiv:1712.09776*.

[37] Menden, Michael P., Dennis Wang, Mike J. Mason, Bence Szalai, Krishna C. Bulusu, Yuanfang Guan, Thomas Yu et al. "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen." *Nature communications* 10, no. 1 (2019): 1-17.

[38] Choobdar, Sarvenaz, Mehmet Ahsen, Jake Crawford, Mattia Tomasoni, David Lamparter, Junyuan Lin, Benjamin Hescott et al. "Open community challenge reveals molecular network modules with key roles in diseases." (2018).

[39] Ahsen, Mehmet Eren, Robert M. Vogel, and Gustavo A. Stolovitzky. "Unsupervised Evaluation and Weighted Aggregation of Ranked Classification Predictions." *Journal of Machine Learning Research* 20, no. 166 (2019): 1-40.

[40] https://www.australianmining.com.au/news/oz-minerals-unearthed-award-1m-prize-for-exploration-contest/ (2019).

[41] 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 (2018).

Figure and Table Legends:

Figure 2: *High-level architecture of the custom-built challenge platform depicting data and model flow during challenge operation following the model-to-data paradigm: challenge participants do 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 behind a secure firewall and then provides back model performance results to participants (orange dashed arrows). This is fundamentally different to conventional crowdsourced challenge setups.*

Figure 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 per 24h and (c) FA/24h plotted against S.*

Figure 3: *In order to label 24h of EEG recordings an unassisted human annotator has to review all 24h of raw EEG data (top). Using the systems developed by teams Otameshi, EpiInsights, Ids_cpmp and Team SG a clinician will only have to review all segments which the system automatically detects, i.e. 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 24h raw EEG data recording is ~0.2% [22] or ~3min. The best models achieve a minimum annotation overhead of 7min which therefore allows to reduce the total amount of raw EEG data to be reviewed by a human annotator from 24h down to 10min 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.*

Figure 4: *An engineering step introducing a hyperparameter which allowed to tune sensitivity vs. false alarm rate was part of 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 which can be used as assistive labelling tools for raw EEG data annotation: instead of having to review the entire raw data, human annotators only have to review the events detected by the system. (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.*

Figure 5: *False alarm rates per 24h plotted against detection sensitivity going from 75% sensitivity level to the maximum achievable sensitivity for each system. Note that team AI4MH is not included since their solution could not reach the 75% sensitivity level by applying the engineering step.*

Figure 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 more than 90% and two-order of magnitude data reduction factors.*

Figure 7: *(a) Applying the engineering step introduced by teams Otameshi and Ids_cpmp allows to bring the models of teams Otameshi, EpiInsights and Team SG up to maximum detection sensitivities of 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 levels the lowest data reduction factor (22, Team SG) still allows to compress 24h of raw EEG data down to a ~1h-short segment of raw EEG data to be reviewed by the human annotator.*

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

Table 2: *Overview of performance parameters achieved by the final valid submitted models of all teams against the blind held-out validation dataset and 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.*

Supplemental Information

I) Challenge timeline:

The IBM Deep Learning Epilepsy Detection Challenge was launched on September 12, 2018 and ran for 24 weeks. A detailed challenge timeline is provided in Figure S1.

Figure S1: *Timeline for the IBM Deep Learning Epilepsy Detection Challenge from launch to completion. During the Competitive Phase participants submitted their models for evaluation on the validation data set. In the Evaluation Phase teams were first submitting their final models during the Submission Period. During the Scoring Period the challenge organizing team validated valid final submissions on the blind test set which had been shielded from participants at all times. Only model performances on the blind test set were used for final scoring.*

II) Procedure for selecting the challenge winner:

At the completion of the challenge, 7 teams submitted valid final solutions as per the challenge rules. Teams were allowed to make multiple submissions during the final submission stage. To determine which team had won the challenge, we chose the single best submission per team out of all the submissions a team had made and evaluated the selected models independently on the blind test set. The model selection and evaluation processes are described here below:

1. State-of-the-art performance of machine learning based automatic seizure detection models applied to the challenge dataset and using the same TAES evaluation metric stands at 30.83% sensitivity at a false alarm rate of 7 FA/24h [36]. While this false alarm rate is approaching that of human experts, the corresponding sensitivity level lacks any clinical relevance. 75% detection sensitivity constitutes the threshold for clinical applicability of an automatic seizure detection system [24]. Therefore, to be eligible for final ranking as per the challenge rules submitted models had to achieve at least 75% detection sensitivity on the blind test set. This was a hard requirement:

submissions achieving lower than 75% sensitivity on the blind test set were not considered for final rankings regardless of their evaluation metric E scores.

2. Models satisfying step 1 were then ranked by their evaluation metric E with the lowest E ranking highest.

Table ST1 shows the performance results of all seven valid final submitted models on the blind test set.

Scoring of best final submissions on blind test set

Team	Evaluation metric	Sensitivity	FA/24hours
Ids_cpmp	1.94	53.25	104.06
Team SG	3.21	65.40	211.08
AI4MH	3.51	14.68	51.54
LateStarters	3.51	45.79	161.31
Otameshi	3.70	57.45	215.83
AP	5.93	45.16	268.33
EpiInsights	7.14	55.64	398.24

Table ST1: *Performance results of all seven valid final submissions as validated on the blind test set. No team reached the 75% sensitivity threshold thus no challenge winner was declared.*

No team reached the required 75% sensitivity mark, and as a result no team was declared the official challenge winner. However, team Otameshi introduced an optional engineering step as part of their final submission (step 3.3 in the process flow chart below) which periodically added synthetic false alarms and included a hyperparameter for tuning sensitivity and false alarm rates. This feature allowed to demonstrate feasibility of a tunable annotation assistant with sensitivities beyond 75%. We further investigated the impact of the engineering step when applied to all valid final submissions and provide the results of this exploration in Figure S2 and Figure S3 below.

III) Process flow for the solution of team Otameshi:

1. Pre-processing:

1.1 While the original problem setting is to predict whether or not a given one second interval corresponds to an ictal or normal state in the raw EEG data, in this solution the label for a given one-sec interval at time T is predicted through an 4-sec segment of EEG data (T-3, T-2, T-1, and T).

1.2 To mitigate the imbalance of true and false labels, the true label data are increased using oversampling by a factor of 1.6667.

1.3 While the EEG data was given in a form of 'FP1-F7;F7-T3;T3-T5;T5-O1;FP2-F8;F8-T4;T4-T6;T6-O2;T3-C3;C3-CZ;CZ-C4;C4-T4;FP1-F3;F3-C3;C3-P3;P3-O1;FP2-F4;F4-C4;C4-

P4;P4-O2', where 20 electrode differentials are computed, in this solution more combinations of electrode differentials are computed to increase accuracy.

1.4 Each 4-sec EEG data segment is divided into four 1-sec intervals, and Fast Fourier Transform is applied to each 1-sec data segment.

2. Model:

2.1 The solution uses a convolutional neural network (a combination of convolution, max pooling, and drop out) for each 2-sec interval of data both for raw EEG data and FFT outputs (note that the 4-sec interval data is divided into four 1-sec intervals during pre-processing).

2.2 A dense network is used to combine the outputs of the neural networks.

3. Post-processing:

3.1 The output of the trained model is averaged. The prediction value at time T is then calculated as an average in the time interval between T-14 and T+14.

3.2 A ictal or normal label is assigned to each 1-sec time interval based on the average value calculated in step 3.11 by using a threshold value of 0.12.

Optional Engineering Step 3.3 An artificial false ictal label is inserted every 170 seconds.

3.4 Remove ictal intervals shorter than 30 seconds.

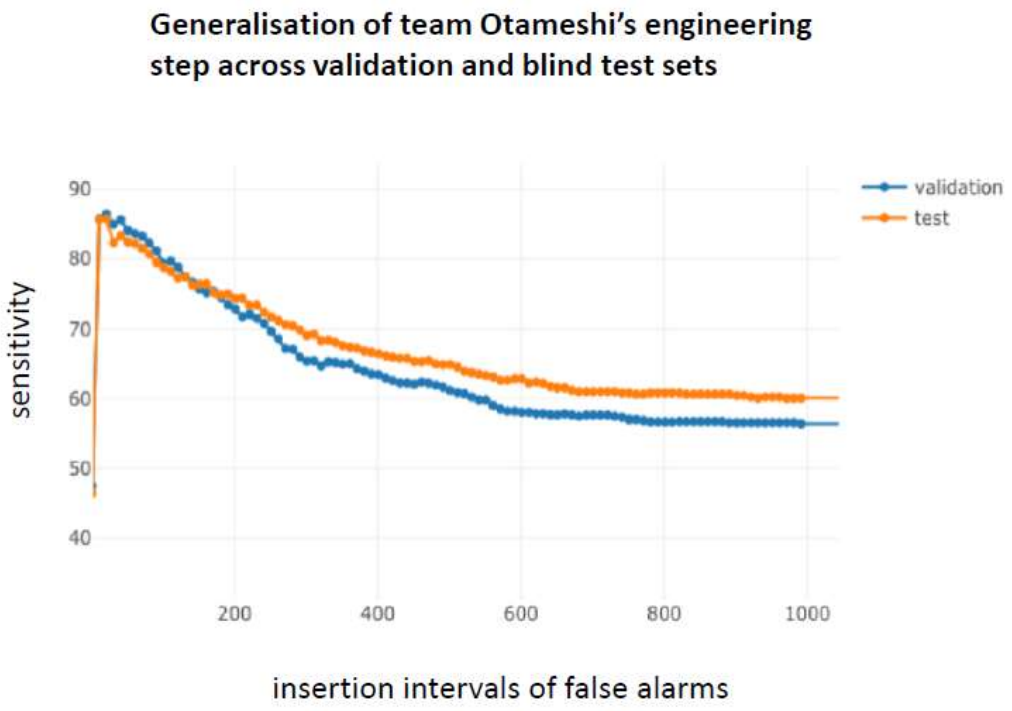

Figure S2: *Variation of insertion intervals of synthetic false alarms affect the detection sensitivity. As demonstrated here for the submission of team Otameshi, the performance of the model generalizes to the blind test set when the hyperparameter of the engineering step is tuned on the validation set.*

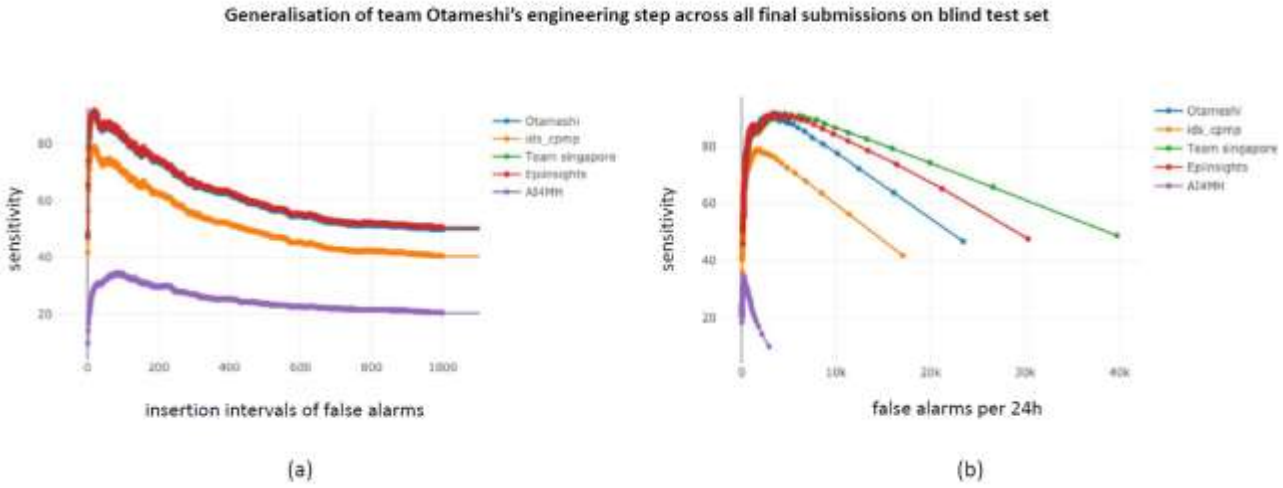

Figure S3: *The engineering step introduced by team Otameshi generalizes across all seven final submissions of the other teams. (a) Insertion intervals of synthetic false alarms affect the detection sensitivity for all submissions following the same trend as shown for team Otameshi in Figure S2. (b) Insertion of synthetic false alarms impacts the number of total false alarms per 24h as well as the detection sensitivity.*