

# Labeler-hot Detection of EEG Epileptic Transients

Lukasz Czekaj<sup>1</sup>, Wojciech Ziembła<sup>1</sup>, Pawel Jezierski<sup>2</sup>, Pawel Swiniarski<sup>1</sup>,  
Anna Kolodziejak<sup>1</sup>, Pawel Ogniewski<sup>1</sup>, Pawel Niedbalski<sup>1</sup>, Anna Jezierska<sup>3,4</sup>, Daniel Wesierski<sup>4</sup>

**Abstract**—Preventing early progression of epilepsy and so the severity of seizures requires effective diagnosis. Epileptic transients indicate the ability to develop seizures but humans overlook such brief events in an electroencephalogram (EEG) what compromises patient treatment. Traditionally, training of the EEG event detection algorithms has relied on ground truth labels, obtained from the consensus of the majority of labelers. In this work, we go beyond labeler consensus on EEG data. Our event descriptor integrates EEG signal features with one-hot encoded labeler category that is a key to improved generalization performance. Notably, boosted decision trees take advantage of singly-labeled but more varied training sets. Our quantitative experiments show the proposed labeler-hot epileptic event detector consistently outperforms a consensus-trained detector and maintains confidence bounds of the detection. The results on our infant EEG recordings suggest datasets can gain higher event variety faster and thus better performance by shifting available human effort from consensus-oriented to separate labeling when labels include both, the event and the labeler category.

## I. INTRODUCTION

Misinterpretation of scalp electroencephalogram (sEEG) is not uncommon in clinical practice [1],[2]. At the same time, it can have severe negative consequences on health and well-being of patients undergoing epileptic diagnosis [3]. Developing algorithms that reliably assist clinicians in EEG inspection is thus an important challenge.

Epilepsy is a chronic disease that affects dozens of millions of people worldwide, being the second neurological disorder after stroke. Nearly 85% of the affected population belongs to developing countries. Roughly 2.4 million new cases of epilepsy occur every year globally. Epilepsy is often a consequence of motor vehicle accidents. As its occurrence increases with age, aging societies are especially at risk to suffering from epilepsy. Patients with epilepsy have a mortality rate significantly higher than that of the general population [4].

Meanwhile, diagnostics of the disease can be time-consuming – from hours to days, is expensive, and requires long clinical experience of the personnel. Gold-standard procedure for diagnosing epilepsy is measuring the electric activity of the cortex with sEEG. The modality uses a lattice of electrodes that are placed along the scalp. Inspection of EEG aims at finding patterns that mark abnormal electric activity of the brain. Among them, transient epileptic patterns indicating tendency toward seizures are of special interest.

The prevalent approach to detection of epileptiform EEG discharges relies on machine learning algorithms that train a decision function in some feature space on an annotated dataset of EEG micro events. Recent validation studies show that human experts continue to outperform algorithms in detection of epileptiform discharges in sEEG [5]. However, annotating pathological events, such as spikes, sharp waves, slow waves, and their complexes, is far from evident. A human expert can confuse pathological with benign events that share similar morphology [6]. Low signal-to-noise ratio and the presence of artifacts are other confounding causes of labeling errors [7].

Datasets are annotated, in effect, by a designated group of hospital personnel, with multiple but noisy labels per event. Ground-truth event labels for training and testing usually are obtained then through majority-voted consensus of labelers [1],[8],[9]. However, the amassed multiple labels show only low-to-moderate inter-rater agreement (IRA) [9]. The majority of neurologists have no neurophysiology fellowship training and there is a substantial discrepancy in event interpretations between board-certified academic clinical neurophysiologists [1]. Moreover, technicians, who are more available than clinicians for annotating EEG [8], often have less clinical experience and qualifications. The features of raters were analyzed in [3] that generally concluded the highest IRA was attributed to board-certified annotators. The groups of features of EEG signal, in turn, were selected and evaluated in [1] that indicated wavelets led to higher IRA.

This work addresses the problem of training an EEG event detector on single and multiple labels per event when labels are provided by imperfect experts. Traditional scenario for training an EEG event detection algorithm has relied on consensus of the majority of labelers that determined ground truth event labels. The multiple labels have only low-to-moderate IRA though. We go beyond labeler consensus on EEG data. We demonstrate that a detector can gain higher recall-precision performance through training on single instead of multiple labels when: (i) more events are sampled across time and recordings thereby increasing variety of training data and (ii) labels identify labelers apart from events. We achieve this by integrating groups of signal features with one-hot encoded labeler category in boosted decision trees training regime. The classifier then selects optimal feature subsets of epileptic events for training the event detector. We show that the proposed labeler-hot features are a key to higher generalization performance of the classifier. To our knowledge, we are the first to train EEG classifiers from consensus-free labels of imperfect experts.

<sup>1</sup>Elmiko Biosignals, Poland

<sup>2</sup>Institute of Psychiatry and Neurology, Poland

<sup>3</sup>Systems Research Institute of the Polish Academy of Sciences, Poland

<sup>4</sup>Gdansk University of Technology, Poland

## II. RELATED WORK

Different approaches to learning from noisy labels have been studied before. Ground truth can be estimated from multiple, noisy labels using crowdsourcing. Besides naive majority voting, more sophisticated algorithms based on EM and labeler reliability estimation were proposed [10], [11], [12] but require high redundancy of labels [13]. Recently, to overcome high redundancy regiment, an EM algorithm used predicated labeled as ground truth to estimate labeler confusion matrix [14]. There are also results specifically in the area of time series labeling, which are more related to EEG annotations than image labeling [15], [16]. Another line of works tweaks loss function to incorporate assumption about uniform noise process disturbing labels [17], [14], [15]. There was significant amount of work in the area of active learning [18], [19] that ask for more labels for inconsistent examples.

Allocation of work (i.e. multiple labeling vs single labeled but larger dataset) was studied in [20], [21] founding that repeated labeling performs better if quality of labelers is below some threshold.

Unlike other approaches, that model labeler quality weights training examples, we focus on modeling individual labeler "styles". Specifically, our approach attempts to predict which labeler says what about given EEG example. To our knowledge, similar approaches were used for the first time in [22] and then generalized as "crowd layer" in [23] and for time series annotations in [24]. Approach presented in [22] is based on learning of logistic regression classifier for each labeler on the features of obtained from Inception-V3. Then single labeler scores are aggregated by weighted averaging. In the training phase, loss function takes into account only the output corresponding to the labeler which provided the example.

Our approach uses XGBoost learning and explicitly encoded labeler as a feature. That differs our approach from [22]. We also evaluate different methods of detection at test time as then no labeler information is provided. Besides averaging, we also test labeler agnostic approach. Moreover, we demonstrate our approach on time series annotations rather than image labeling.

Detection of epileptiform EEG discharges has been addressed before as well. Detection has usually used time domain features (e.g. amplitude, duration, curvature/sharpness, complication, fractal dimension, sample entropy), frequency domain features (e.g. spectral power density, Hjorth's parameters, phase congruence) or wavelet domain features (e.g. wavelet coefficients). Different classification algorithms were applied: SVM, logistic regression, boosted trees, random forests. There were also experiments with clustering and anomaly detection methods and dynamic time warping. For comprehensive review see [25], [26]. Spike detection brought attention in deep learning community [27]. Commercial implementations also exist [8].

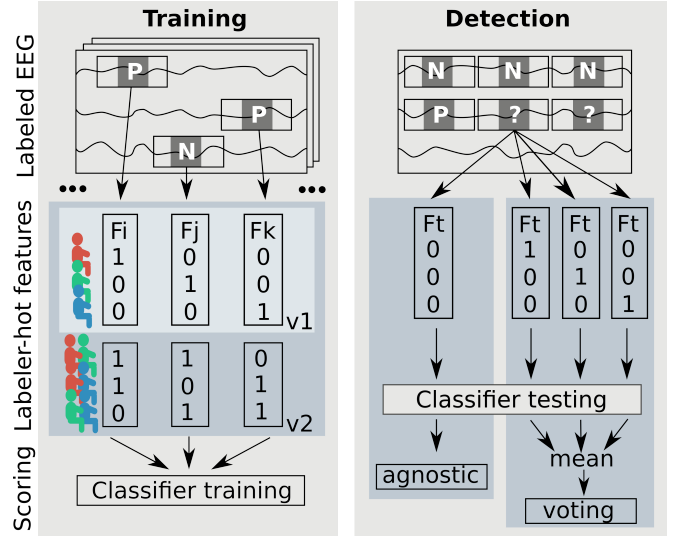


Fig. 1. Labeler-hot detection of EEG events.

## III. METHOD

We consider that task of detecting EEG micro-events in a single channel irrespective of other channels – our detector processes each channel separately. The flowchart of our method is depicted in Fig. 1.

### A. Signal description

Our descriptor is composed of three windows, a central window of 0.2 sec. duration and two neighbourhood windows of 0.8 sec. duration each (Fig. 2). Then, we calculate the following features of the windowed EEG signal and stack them column-wise into a descriptor:

- Time series anomaly score, i.e. linear model prediction error (refer to Fig. 2 for details);
- FFT features (log power for frequency) in central window, neighbourhood and quotient of window and neighbourhood features;
- Teager Energy for central window, neighbourhood and their quotient;
- Quotient of waveform length for window and neighbourhood;
- Standardised statistics (mean, standard deviation, skewness, min, max) of continuous wavelet (Ricker wavelet) transform coefficients for detection window; we use signal standardization according to the neighbourhood;
- Statistics (mean, standard deviation, skewness, min, max) of EEG signal difference with lag 1 in the central window.

### B. Learning

The experience of labelers manifests itself in specific expert annotation styles that can be learned by the event detection algorithms. To this end, we propose to integrate the signal descriptor with the one-hot encoded labeler category. We explore 2 variants of such an encoding:

- 1) single expert category (v1) – each expert corresponds to one row in the descriptor; we set to 1 the row of

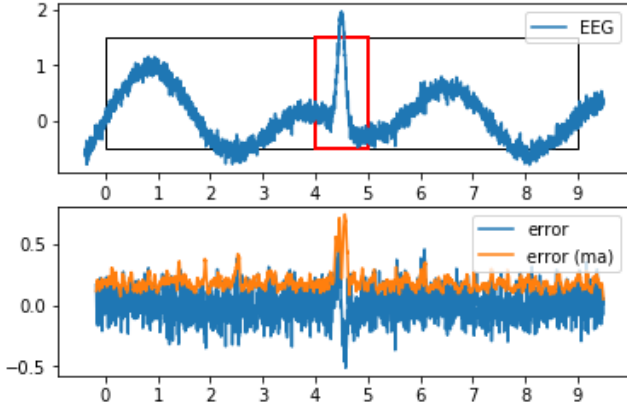


Fig. 2. *Signal description (upper panel)* - we consider example localized at  $t = 4.5$  in the central window (red box) and classify it into positive or negative class. Signal descriptor also uses the neighbourhood of the central window (the black boxes on the left and right of central window); *Time series anomaly score (lower panel)* - first we sub-sample the original EEG (factor 3), then train auto-regressive model of order 5 on the neighbourhood. Finally, we use model to predict detection in the central window. Prediction error increases if there is an epileptic transient in the window.

the expert who annotated a given example, otherwise the row is 0,

- 2) pair of experts category (v2) through one-hot encoding of single experts and one-hot encoding of 2-expert groups – the same rows as in v1 and one row for every combination of a pair of experts; we set to 1 the rows that correspond to groups that contain the expert who annotated a given example, otherwise the row is 0.

We then use an XGBoost classifier for training our decision function. The features that describe the signal and the labels are input together with a class label to the XGBoost classifier. The classifier outputs predictions as probability that a tested example is positive.

### C. Detection

During training, we have information about who labeled what but for test examples we lack such cues. Moreover, we argue it is better not to predict annotations from specific experts but to make predictions that can better correspond with the ground truth. To this end, we propose two detection methods:

- 1) expert agnostic – all rows of descriptor related to labels are set to 0, what neglects labeler style,
- 2) expert voting – each example is considered to have been individually annotated by each expert and thus is processed as a set of (v1) descriptors. Then, the prediction outputs are mean-averaged as a final result.

## IV. EXPERIMENTS AND RESULTS

### A. Dataset and evaluation criteria

Our dataset (Tab. I) consists of 30 EEG recordings of cohort of infants with tuberous sclerosis. The dataset is split into 24 training (18 patients) and 6 test (6 patients) recordings. The age of patients spans 3-14 months in the

labeler	training dataset			test dataset		
	recording	pos. #	neg. #	recording	pos. #	neg. #
L1	R7–30	3118	542K	R1–6	1857	729K
L2	R7–30	498	546K	R1,R3–6	374	542K
L3	R8–14	1199	380K	R1,R3–6	981	545K
	R16–21					
	R23					
L4	R26–28					
L5	–	–	–	R1,R2	270	357K
L6	–	–	–	R2	342	196K
L7	R7,R15	288	163K	R2	347	195K
	R22,R24					
	R25,R29					
L7	R30					
L7	–	–	–	R4	616	90K

TABLE I

SUMMARY OF OUR EEG DATASET OF EPILEPTIC TRANSIENTS IN INFANTS WITH TUBEROUS SCLEROSIS ( $K = \times 10^3$ ).

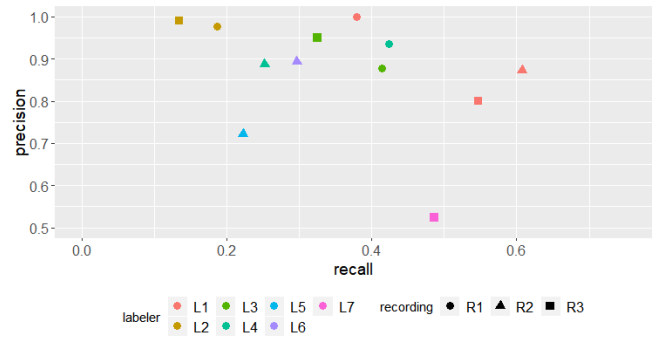


Fig. 3. *Labeler quality* - we show precision and recall of labels. Each labeler is compared against ground truth, obtained from majority voting of 3 other labelers. Results are presented for 3 test recordings. There were more labels in testing dataset than in the training dataset. Labelers, who were present in training set, are: L1, L2, L3, L6.

training and 2-26 months in the test recordings. Each  $\sim 1$ h long recording contains 5 blocks of 18 channels of dense annotations in bipolar Banan 2 montage (20-10 standard). From 3 to 4 EEG technicians with years of experience were asked to segment each channel from the same blocks of 5 sec. into adjacent fragments of max. duration of 2 sec. Each fragment was labeled independently by each expert either as: (N) artifact, slow wave, sleep spindle, norm, other, (P) sharp wave, spike, sharp wave complex, or spike complex. In our binary classification setting, the first group is negative (N) and the second group is positive (P). Experts who annotated training data also annotated test data. All recordings and annotations were acquired with Elmiko EEGDigiTrack hardware and software.

**Evaluation metric** The ground truth that serves to evaluate the trained detectors was acquired using majority voting from a set of 3 experts. We prepare 5 subsets of negative test examples by randomly sampling the whole negative testset. As annotations can be longer than 0.2s, we obtained test dataset by sampling 0.2s example windows from the consensus windows. The positive testset is fixed and has 4223 positive examples. Each pair of negative and positive testsets is imbalanced, where the count of negative to positive

examples is 20 : 1 thereby reflecting prevalence of negative events in EEG.

We used average precision (AP) score from the precision-recall curve to evaluate our method. If there are more than 3 experts for given recording, we have several consensus annotations. First we average the AP scores inside recording (in case there are multiple consensus annotations) and then average the AP scores of the recordings. In this way each recording has the same impact on the final score.

**Expert vs consensus** In order to assess the quality of the labelers, we compare their individual performance with the consensus-based ground truth from the rest of labelers on the test data in Fig. 3. The labelers tend to have higher precision than recall. Lower recall indicates that experts are likely to miss epileptic events in the electroencephalogram.

### B. Scenarios for learning from noisy labels

We describe our scenarios (see Fig. 4) for work allocation that address the problem of training an EEG event detector on labels from multiple but imperfect experts. Notably, assuming budget, availability, and time constraints, we are interested whether a group of medical labelers should annotate (i) the same recordings at the same time instants (A,B), (ii) the same recordings but at different time instants (C), or (iii) different recordings (D) in order to build a dataset that will allow training the best event detector.

The A-scenario refers to obtaining ground truth labels from the majority-voted consensus of labelers and thus requires multiple labels per event. The B-scenario uses the same training examples as in A but replicates each one  $K$  times and augments the replica with the corresponding category of  $K$  labelers. Hence, scenario A has  $K \times$  fewer training data than do the scenarios B, C, and D, which have single labels per event. Importantly though, the group of experts perform the same amount of work in each scenario.

**Sampling** We are given  $N = 24$  recordings in the training dataset, multiply annotated by  $K = 3$  labelers. We randomly sample 5 times either (i) 8 out of 24 recordings (A,B,C) or (ii) a disjoint assignment of 3 labelers to 24 recordings (D). Then, we sample 5 times 100 training examples of positive and negative class of events on the time axis per each sampled recording. In total, we have 25 different realisations of training data in the form of (recording, labeler)-pairs in each scenario with 800 and 2400 positive and negative examples for A and B,C,D scenarios, respectively.

### C. Quantitative results

The proposed detection methods (sec. III-C) are evaluated in Fig. 5. The voting method, where all labeler-hot encoded classifier responses equally contribute to the final score, is consistently better than the consensus-based method as well as the agnostic-based detection method that adjusts its observations to no labelers. We show though our next results using the agnostic method.

We show the dependence of the AP scores on the size of the training dataset (i.e. size per recording and labeler) in Fig. 7. The results were obtained for the C-scenario with

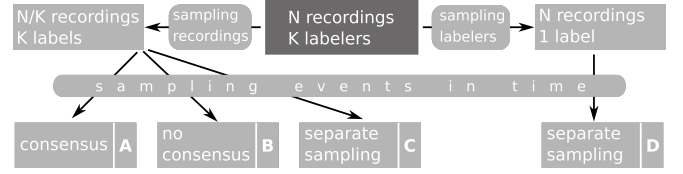


Fig. 4. Scenarios A–D for training EEG event detectors from noisy labels.

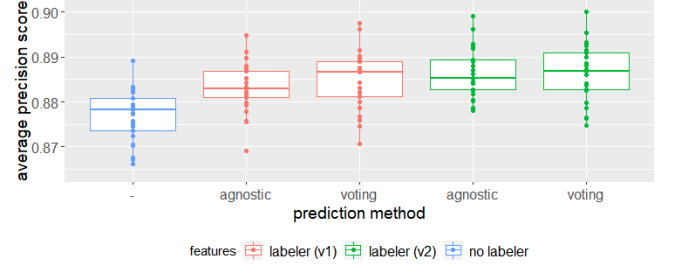


Fig. 5. Detection methods - box plots compare two detection methods on test data: labeler agnostic and voting (sec. III-C). Each point depict average precision score for single experiment. Voting performs systematically better than labeler agnostic method for both types of labeler-hot descriptors.

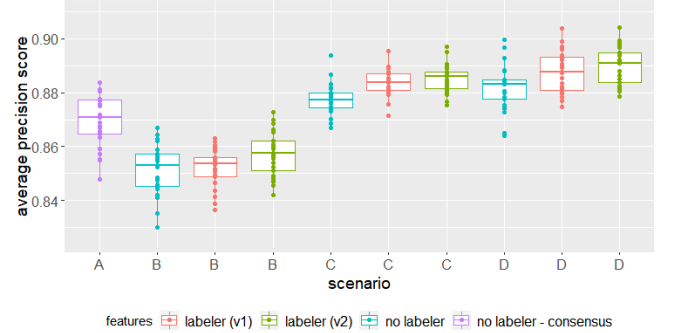


Fig. 6. Comparison of learning scenarios - box plots of average precision scores for scenarios A,B,C,D (sec. IV-B), descriptors v1,v2 (III-B), and agnostic detection method (cf Fig. 5). Each point represents a single experiment. The B-scenario has poorest performance. We observe including labeler category into the signal descriptor (C,D-scenarios) leads to systematically better results and maintains the confidence bounds of the detection wrt to consensus-based detector (A-scenario).

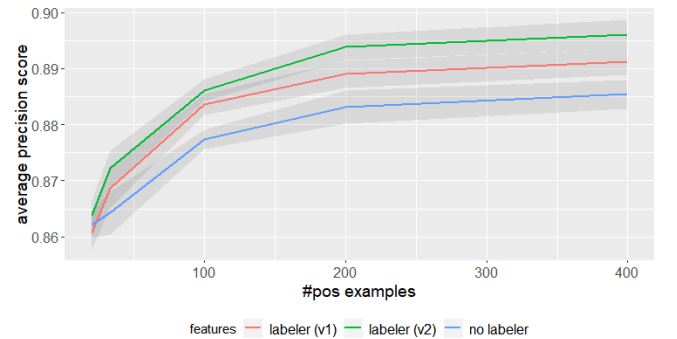


Fig. 7. Detection performance wrt increasing number of training examples We present median and confidence intervals of average precision score for experiments. Descriptors enhanced with labeler category perform systematically better than the consensus-based detector.

v1 and v2 descriptors (III-B) and the agnostic detection method (cf Fig. 5). We observe the performance of the proposed labeler-hot detector outperforms the consensus-based detector despite inflating the training dataset. Our labeler-hot detector systematically takes advantage of the growing number of training data as well as of the descriptors that are enhanced with labeler category.

We evaluate the scenarios from sec. IV-B in Fig. 6. We find that allocating labelers to disjoint annotations over time axis (C) and over recordings (D) improves performance over collective annotations (A) up to  $\sim 2\%$  in the AP score. Including the labeler category helps in every scenario (B,C,D). However, the B-scenario is always worse than the A-scenario. In the proposed experimental setting, this indicates that the detector performs best when variety of training data is increased without the need for increasing allocated work and a descriptor includes labeler category.

## V. CONCLUSIONS

We describe an effective approach to leveraging individual expertise of medical labelers. Experts have unique strengths in annotating specific EEG data – some experts might feel more comfortable with annotating artifacts while others with annotating spikes. Such expert preferences manifest themselves in specific annotation styles that can be learned by the event detection algorithms. To this end, our approach integrates the signal descriptor with variants of one-hot encoded labeler categories and shifts available human effort from consensus-oriented to separate labeling thereby increasing the variety of the training dataset. Both propositions are a key to increased performance of EEG event detectors.

## ACKNOWLEDGMENT

This work was financed in part by the Epimarker project under agreement STRATEGMED3/306306/4/NCBR/2017.

## REFERENCES

- [1] Elham Bagheri, Justin Dauwels, Brian C Dean, Chad G Waters, M Brandon Westover, and Jonathan J Halford. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology*, 128(10):1994–2005, 2017.
- [2] Nese Dericioglu and Pinar Ozdemir. The success rate of neurology residents in eeg interpretation after formal training. *Clinical EEG and neuroscience*, 49(2):136–140, 2018.
- [3] Jonathan J Halford, Amir Arain, Giridhar P Kalamangalam, Suzette M LaRoche, Bonilha Leonardo, Maysaa Basha, Nabil J Azar, Ekrem Kutluay, Gabriel U Martz, Wolf J Bethany, et al. Characteristics of eeg interpreters associated with higher interrater agreement. *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, 34(2):168, 2017.
- [4] Samden D Lhatoo and Josemir WAS Sander. Cause-specific mortality in epilepsy. *Epilepsia*, 46:36–39, 2005.
- [5] Jonathan J Halford, M Brandon Westover, Suzette M LaRoche, Micheal P Macken, Ekrem Kutluay, Jonathan C Edwards, Leonardo Bonilha, Giridhar P Kalamangalam, Kan Ding, Jennifer L Hopp, et al. Interictal epileptiform discharge detection in eeg in different practice settings. *Journal of Clinical Neurophysiology*, 35(5):375–380, 2018.
- [6] Balagopal Santoshkumar, Jaron JR Chong, Warren T Blume, Richard S McLachlan, G Bryan Young, David C Diosy, Jorge G Burneo, and Seyed M Mirsattari. Prevalence of benign epileptiform variants. *Clinical Neurophysiology*, 120(5):856–861, 2009.
- [7] Jonathan J Halford. Computerized epileptiform transient detection in the scalp electroencephalogram: Obstacles to progress and the example of computerized eeg interpretation. *Clinical Neurophysiology*, 120(11):1909–1915, 2009.
- [8] Mark L Scheuer, Anto Bagic, and Scott B Wilson. Spike detection: Inter-reader agreement and a statistical turing test on a large data set. *Clinical Neurophysiology*, 128(1):243–250, 2017.
- [9] Jonathan J Halford, Robert J Schalkoff, Jing Zhou, Selim R Benbadis, William O Tatum, Robert P Turner, Saurabh R Sinha, Nathan B Fountain, Amir Arain, Paul B Pritchard, et al. Standardized database development for eeg epileptiform transient detection: Eegnet scoring system and machine learning analysis. *Journal of neuroscience methods*, 212(2):308–316, 2013.
- [10] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.
- [11] Denny Zhou Yuchen Zhang, Xi Chen and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *NIPS*, pages 1260–1268, 2014.
- [12] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Gaussian process classification and active learning with multiple annotators. In *International Conference on Machine Learning*, pages 433–441, 2014.
- [13] et al. Jonathan Bragg, Daniel S Weld. Optimal testing for crowd workers. *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pages 966–974, 2016.
- [14] Zachary C. Lipton Ashish Khetan and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint*, 2018.
- [15] Rahul Gupta, Kartik Audhkhasi, Zach Jacokes, Agata Rozga, and Shrikanth Narayanan. Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach. *IEEE Transactions on Affective Computing*, 9:76–89, 2016.
- [16] Chen Wang, Phil Lopes, Thierry Pun, and Guillaume Chanel. Towards a better gold standard: Denoising and modelling continuous emotion annotations based on feature agglomeration and outlier regularisation. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, AVEC’18*, pages 73–81, New York, NY, USA, 2018. ACM.
- [17] Pradeep K Ravikumar Nagarajan Natarajan, Inderjit S Dhillon and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, pages 1196–1204, 2013.
- [18] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, pages 2035–2043, 2009.
- [19] Ashish Khetan and Sewoong Oh. Achieving budget-optimality with adaptive schemes in crowd-sourcing. *Advances in Neural Information Processing Systems*, pages 4844–4852, 2016.
- [20] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622, 2008.
- [21] Victor S Sheng Panagiotis G Ipeirotis, Foster Provost and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, pages 402–441, 2014.
- [22] Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*, 2018.
- [23] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. *AAAI Conference on Artificial Intelligence*, 2018.
- [24] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. Deep learning for continuous multiple time series annotations. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, AVEC’18*, pages 91–98, New York, NY, USA, 2018. ACM.
- [25] Scott B Wilson and Ronald Emerson. Spike detection: a review and comparison of algorithms. *Clinical Neurophysiology*, 113(12):1873–1881, 2002.
- [26] Fathi E. Abd El-Samie, Turkey N. Alotaiby, Muhammad Imran Khalid, Saleh A. Alshebeili, and Saeed Abdullah Aldosari. A review of EEG and MEG epileptic spike detection algorithms. *IEEE Access*, 6, 2018.
- [27] Alexander Rosenberg et al. Epileptiform spike detection via convolutional neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 754–758, 2016.