KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Zhang, J., Chatzichristos, C., Vandecasteele, K., Swinnen, L., Broux, V., Cleeren, E., Van Paesschen, W. and De Vos, M. (2022), **Automatic annotation correction for wearable EEG based epileptic seizure detection Journal of neural engineering, vol. 19, Feb. 2022, 1-13..**
Archived version	Author manuscript: pre-print version
Published version	insert link to the published version of your paper https://doi.org/10.1088/1741-2552/ac54c1
Journal homepage	insert link to the journal homepage of your paper https://iopscience.iop.org/journal/1741-2552.
Author contact	your email jingwei.zhang@esat.kuleuven.be your phone number
Abstract	
IR	url in Lirias https://lirias2.kuleuven.be/viewobject.html?id=3692881&cid=1

(article begins on next page)

Automatic annotation correction for wearable EEG based epileptic seizure detection

Jingwei Zhang[1], Christos Chatzichristos[1], Kaat Vandecasteele[1], Lauren Swinnen[2], Victoria Broux[3], Evy Cleeren[3], Wim Van Paesschen[2,3], Maarten De Vos[1,4]

[1] Department of Electrical Engineering, STADIUS, KU Leuven, Leuven, Belgium
[2] Laboratory for Epilepsy Research, KU Leuven, Leuven, Belgium
[3] Reference center for refractory epilepsy, Department of Neurology, UZ Leuven, Leuven, Belgium
[4] Department of Development and Regeneration, KU Leuven, Leuven, Belgium

Email: jingwei.zhang@esat.kuleuven.be

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Objective. Video-electroencephalography (vEEG), which defines the ground truth for the detection of epileptic seizures, is inadequate for long-term home monitoring. Thanks to their advantages in comfort and unobtrusiveness, wearable EEG devices have been suggested as a solution for home monitoring. However, one of the challenges in data-driven automated seizure detection with wearable EEG data is to have reliable seizure annotations. Seizure annotations on the gold-standard 25-channel vEEG recordings may not be optimal to delineate seizure activity on the concomitantly recorded wearable EEG, due to artifacts or absence of ictal activity on the limited set of electrodes of the wearable EEG. This paper aims to develop an automatic approach to correct the imperfect annotations of seizure activity on wearable EEG, which can be used to train seizure detection algorithms. *Approach*. This paper first investigates the effectiveness of correcting the seizure annotations for the training set with a visual annotation correction. Then a novel approach has been proposed to automatically remove non-seizure data from wearable EEG in epochs annotated as seizures in gold-standard video-EEG recordings. The performance of the automatic annotation correction approach was evaluated by comparing the seizure detection models trained with 1. original vEEG seizure annotations, 2. visually corrected seizure annotations, and 3. automatically corrected seizure annotations. *Results*. The automatic seizure detection approach trained with automatically corrected seizure annotations was more sensitive and had fewer false-positive detections compared to the approach trained with visually corrected seizure annotations, and the approach trained with the original seizure annotations from gold-standard vEEG. *Significance*. The wearable EEG seizure detection approach performs better when trained with automatic seizure annotation correction.

Keywords: epilepsy, automated seizure detection, wearable EEG, automatic annotation correction

1. Introduction

Epilepsy, a common neurological disorder caused by the cerebral neurons' abnormal firing, is affecting about 1% of the world population [1]. Despite the fact that antiepileptic drugs may help control the seizures in many patients, about 30% of patients are still suffering from drug-resistant epilepsy (DRE) and its unpredictable seizures [2]. Typically, both the diagnosis and treatment plans of epilepsy are based on the detection and analysis of epileptic seizures. The current gold-standard method in the detection of epileptic seizures is video-electroencephalography (vEEG) monitoring, which is based on an analysis of the full scalp electroencephalography (EEG) recording that allows the objective quantification and documentation of the seizures [3; 4]. However, vEEG is only available in tertiary or specialty hospitals, which limits its advantages. Its discomfort also makes it insufferable and unsuitable for long-term monitoring [4].

 © xxxx IOP Publishing Ltd

An alternative solution is to record the EEG signals on behind-the-ear (BTE) channels with a wearable device and consequently improve measurement comfort [5-7]. Based on this idea, novel wearable devices have been developed to empower ultra-long-term and unobtrusive EEG monitoring in the home setting [8]. As a considerable amount of data will be collected by wearable devices, automated seizure detection approaches that range from threshold based methods to machine learning based methods become sought-after to replace the tedious visual analysis [9-11]. The central idea of automated seizure detection is to automatically classify sparse segments of rhythmic seizures among a large amount of background EEG.

Currently, most automated seizure detection approaches are based on supervised learning and therefore require an excessive amount of annotated data [12]. However, it is extremely expensive and difficult to collect reliable annotations [13], especially when the data is collected from non-standard EEG montages (e.g. BTE EEG and around-the-ear EEG) [14]. Medical experts may have difficulties in annotating the data from non-standard EEG montages due to the lack of relevant training and knowledge of those non-standard montages [15]. Our previous research in [16] has confirmed the difficulty in annotating the seizures on the BTE channels.

To obtain seizure annotations on wearable EEG data, BTE EEG has currently been recorded simultaneously with full scalp EEG so that annotations can be annotated on the full scalp EEG but could be used for the BTE channel EEG recorded by the wearable device. However, the fact that the seizure is visible on one of the full scalp EEG channels does not guarantee that it is visible on the BTE channels. Seizure patterns can be much lower in amplitude on the BTE channels which makes them easily submerged with artifacts. This potential lack of accuracy in the annotated onset and ending may lead to the deviation from the gold standard annotations. This 'mismatch' implies that the annotations become imperfect as annotations for wearable EEG. Training with imperfect annotations may significantly impact the performance of seizure detection models [17-20].

However, manually correcting those imperfect annotations is tedious and impractical, especially when the annotations have to be annotated by experts. Collecting ground truth annotations from a large number of experts will definitely cost considerable financial and logistical resources [21; 22]. Hence, visually re-annotating is not always a viable option.

Due to the lack of perfect annotations from wearable EEG, training with imperfect annotations from vEEG is an option for wearable-device-based automated seizure detection approaches. Hence, automatic approaches that can effectively learn from so-called imperfect annotations can have an impact on the development of automated seizure detection approaches.

In this paper, we first evaluate our previous seizure detection approach [16], which was trained on a dataset containing BTE channels, on a fully independent dataset with a standalone wearable device. Then, we investigate the improvement in seizure detection performance of correcting the annotations for the BTE channels by a visual correction. Finally, we introduce a novel approach that can automatically differentiate between reliable and potentially incorrect annotations, aiming to replace and improve the 'expensive' and time-consuming visual annotation correction. Comparisons have been conducted between the automatically corrected annotations, the visually corrected annotations, and the full scalp EEG based annotations to show the advantages of corrected annotations. A robust seizure detection framework that could learn with imperfect annotations has been established by including the automatic annotation correction approach in the training process.

2. Data acquisition and visual annotation correction

Two datasets were collected in two projects. The first one, where BTE channels were recorded using a vEEG equipment, was used for algorithm development and its automated seizure detection results were already reported in [16]. The second one,

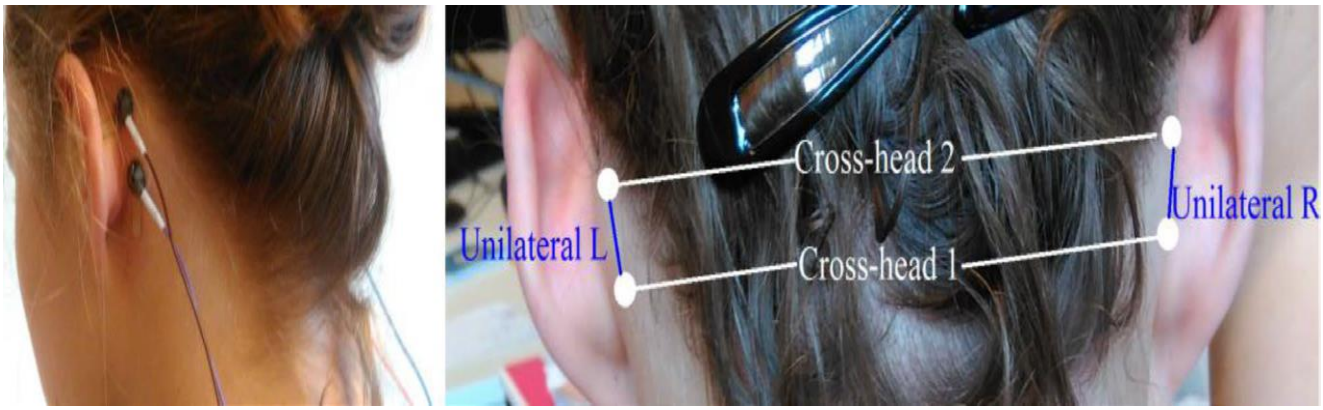


Figure 1. Electrodes setup of behind-the-ear channels used in SeizeIT1. Reproduced with permission from Gu et al [5].

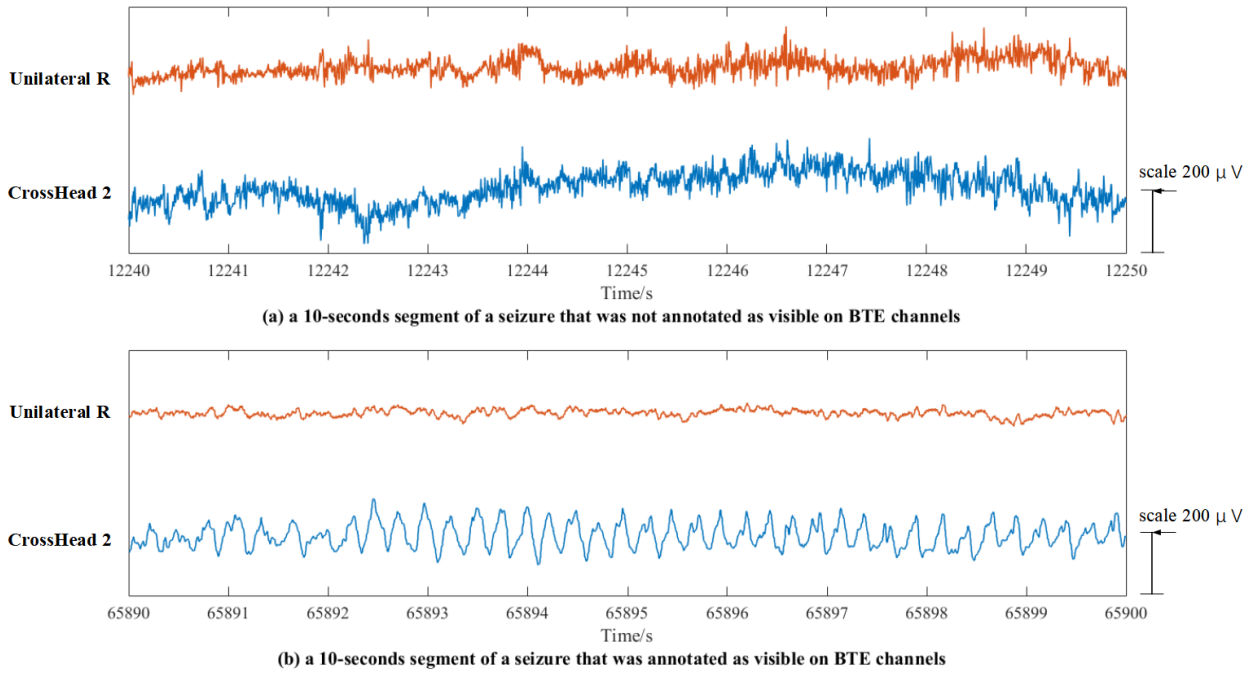


Figure 2. Illustration of seizure annotation in BTE channels. (a) shows a 10-seconds segment of a seizure that showed ictal patterns on the full scalp EEG but no seizure activity can be recognized on the BTE channels and (b) shows a 10-seconds segment of a seizure with visible ictal activity on the BTE channels.

where BTE channels were measured by a fully wearable and wireless device (i.e. the Sensor Dot, Byteflies) [23; 24], was used to validate the performance of the various seizure detection approaches on data from a fully wearable device and the results are reported in this paper.

2.1 SeizeIT1

2.1.1 Data acquisition

The first dataset will be referred to as SeizeIT1 and has been fully described in [16]. EEG signals were recorded with the traditional 10-20 full scalp EEG system using Ag/AgCI electrodes at a sampling rate of 250 Hz. As shown in Figure 1, four additional BTE channels were simultaneously recorded with four additional electrodes located behind the ears (each side has two electrodes). Both full scalp EEG and BTE channel EEG were recorded by the vEEG equipment. As the two crosshead channels were similar in signal recording, only crosshead 2 was kept. Hence, only three channels, i.e. unilateral R, unilateral L, and crosshead 2, have been selected in this dataset. The dataset collected 4034 hours of EEG in total from 42 patients. This dataset was used for all training purposes.

2.1.2 Visual annotation based on full scalp EEG

All the EEG recordings were annotated a neurologist-epileptologist (W.V.P) following standard practice. This dataset contained 182 seizures, 89% of which were focal impaired awareness. The duration of seizures ranged from 11 to 695 seconds with a median of 50 seconds.

2.1.3 Visual annotation based on BTE Channels — an annotation correction

To improve the quality of the annotations for the seizure detection on BTE channels, BTE channels were also blindly annotated by our neurologist-epileptologist (W.V.P) and eningeer (K.V). In this visual annotation correction experiment, 182 BTE-EEG segments had been created; each segments includes a seizure and also the EEG signal before and after that seizure.

The neurologist-epileptologist (W.V.P) together with a eningeer (K.V), who were blinded to the original annotations, selected the segments that have clear ictal EEG pattern and annotated the exact onset and end time of the seizures from those BTE EEG segments. The onset time was determined by the start of the clear ictal EEG pattern, which could be either an increase in amplitude or a change in the frequency compared with the background EEG. For each segment, only the sub-segments that had a clear ictal EEG pattern in the BTE channels were marked as seizures. Other parts of the original seizure segments, which were highly affected by artifacts or did not contain any typical ictal patterns, were not marked as seizures and were not used for the training of the machine learning model in the automated seizure detection approach. Figure 2 shows two 10-second EEG segments that belong to a

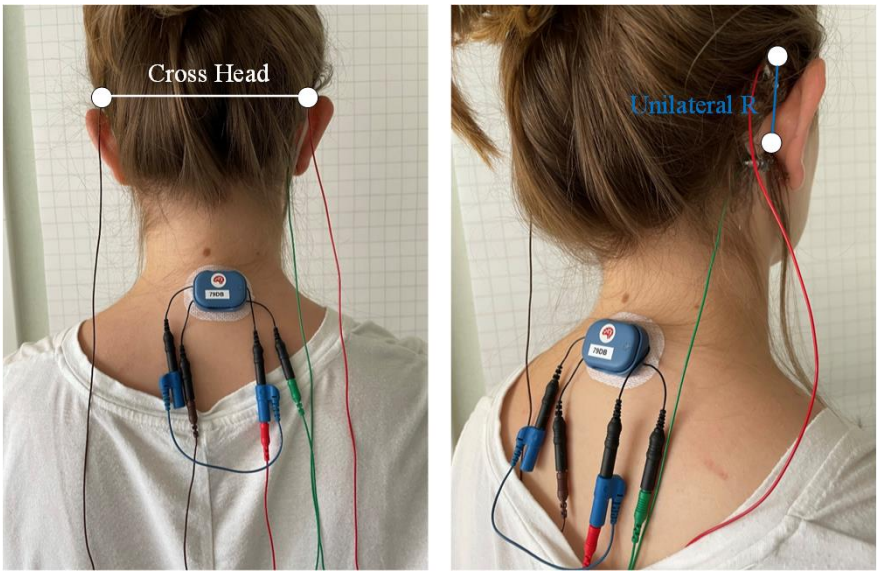


Figure 3. Sensor dot device and its electrodes setup used in SeizeIT2. In this subject, unilateral channel was placed on the right.

seizure that was not annotated as visible on the BTE channels and a seizure that was visible on the BTE channels respectively.

With the visual annotation correction, 66 out of 182 seizures were considered as invisible on the BTE channels; the other seizures were pruned to remove their invisible sub-segments. A total of 7296 seconds of seizure segments were discarded in the visual annotation correction process, which is around half of the total seizures' duration.

2.2 SeizeIT2

2.2.1 Data acquisition

The second dataset will be referred to as SeizeIT2. EEG signals were recorded with an extended version of the standard 10-20 system, according to IFCN guidelines [25] and two additional BTE channels—one crosshead channel and one unilateral channel, placed at the side of suspected epilepsy. The full scalp EEG was still recorded with the vEEG equipment, whereas the BTE channels were recorded by a wearable device, i.e., the Sensor Dot (Byteflies, Belgium) [23; 24], which is shown in Figure 3. Both vEEG recordings and Sensor Dot recordings were sampled at 250 Hz; but some Sensor Dot recordings were longer than their corresponding vEEG recordings due to the early stop of vEEG equipment. After the processing, 3199.9 hours of data from 39 patients were collected in this dataset. Among those patients, the longest monitoring duration was 117.3 hours, and the shortest was 18.6 hours.

2.2.2 Visual annotation based on full scalp EEG

The seizures were also annotated by a neurologist-epileptologist (W.V.P). Seventeen patients had no seizures detected, while the other 22 had seizures ranging from 1 to 13. In total, 67 seizures were included in this dataset, 13 of them were focal aware while 49 of them were focal impaired awareness seizures. The duration of seizures ranged from 10 to 216 seconds with a median of 57 seconds. This dataset was used for testing and was at no point used for training purposes. As there were only annotations made on full scalp in this test set, the seizure detection algorithm was actually asked to detect all the seizures in the test process even if some of them might not be totally visible on the BTE channels.

3. Method

The proposed automated seizure detection framework is shown in Figure 4. Our seizure detection algorithm, which was developed on the SeizeIT1 [16] still forms the core. It is summarized in subsection 3.1. This paper improved the training process by removing the incorrectly annotated samples from the training set. The automatic annotation correction approach allowing the seizure detection approach to learn with the imperfect annotations has been described in subsection 3.2.

3.1 Seizure detection approach

3.1.1 Preprocessing and segmentation

In the preprocessing step, all the BTE channels were bandpass filtered with cutoff frequency at 1 Hz and 25 Hz respectively. The EEG recordings were then split into 2-second segments with 50% overlap between each segment.

3.1.2 Feature extraction

Eighteen features were extracted from each channel to represent the time-domain and frequency domain characteristics [26]. Then three additional entropy features

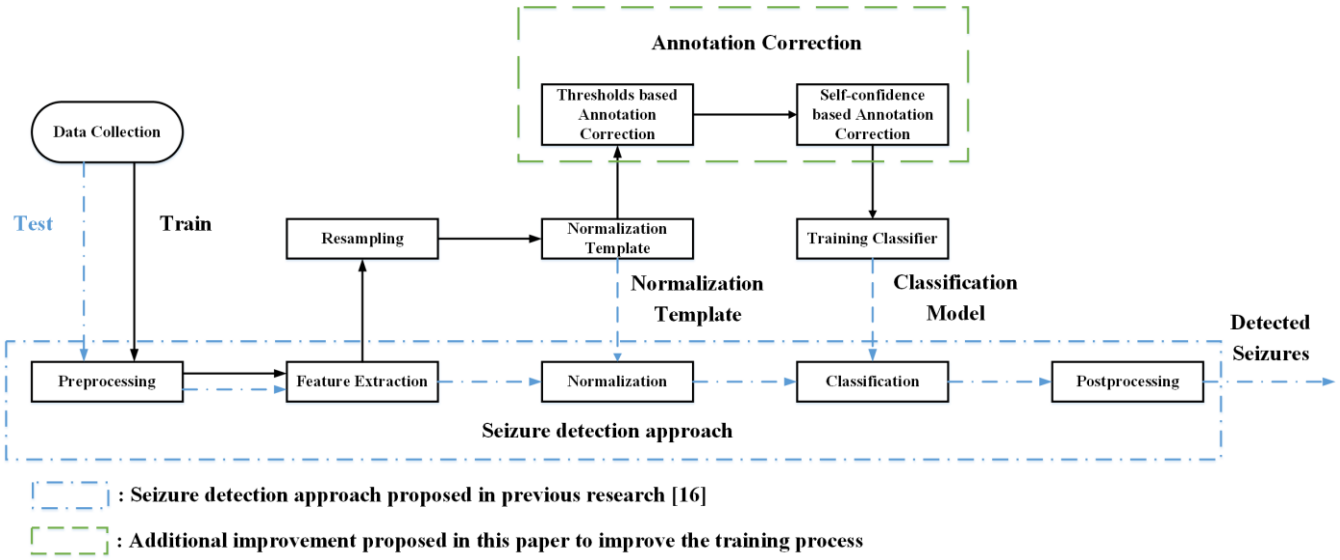


Figure 4. Diagram of the proposed seizure detection framework

were calculated to represent the information-theory-based characteristics [27]. All the features are listed in Table 1. From each BTE channel, 21 features has been extracted independently.

3.1.3 Normalization

After the feature extraction, a normalization was applied on the EEG feature-samples of the training set. The normalization is based on equation (1), where μ denotes the average of each feature and σ denotes the standard deviation of each feature.

$$x_{normalized} = \frac{x_{original} - \mu}{\sigma} \quad (1)$$

Table 1. List of extracted features

Time domain features	1-3. Number of zero crossings, maxima, and minima
	4. Skewness
	5. Kurtosis
	6. Root mean square amplitude
Frequency domain features	7. Total power
	8. Peak frequency
	9-16. Mean and normalized power in frequency bands:
	delta (1-3 Hz)
	theta (4-8 Hz)
	alpha (9-13 Hz),
	beta (14-20 Hz)
	17-18. Mean and normalized power in high-frequency band (40-80 Hz)
Entropy feaures	19. Sample entropy
	20. Shannon entropy
	21. Spectral entropy

The same normalization procedure was followed also for test set feature-samples with the normalizing template calculated based on the training set.

3.1.4 Support vector machine classifier

A non-linear support vector machine (SVM) with a Gaussian kernel was used as the basic classifier. SVM is a supervised learning method that is trained with annotated data. Its optimization objective is to find a hyperplane that maximizes the margin between the given two classes, which could be written as

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} v_i\xi_i \quad (2)$$
$$\text{s.t. } y_i(w^T\varphi(x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0, i = 1,2,\dots,N$$

where (w, b) defines the hypothesis of the classifier; C is the penalty factor and ξ_i is the slack variable. v_i is determined by the number of training samples in different classes:

$$v_i = \frac{N}{2N_+} \; if \; y_i = +1$$

$$v_i = \frac{N}{2N_-} \; if \; y_i = -1$$

SVM is a supervised method that requires the input of training labels, any possible inaccuracy in the labels may lead to a different classification model. Hence, it is important to remove any samples that have incorrect labels from the training set. The classifier will be trained on a balanced dataset, which includes all the seizure samples and a resampled subset of the non-seizure samples. The hyper-parameters of SVM will be optimized by internal cross-validation on the training set.

3.1.5 Post-processing

The feature 'root mean square amplitude (RMSA)' was also used in the post-processing; any samples with an RMSA larger than 150 μV or less than 13 μV were set as non-seizure samples by default. As all the seizures were supposed to last more than 10 seconds, a seizure detection would only be generated when 10 consecutive windows were classified as seizures. However, to allow a few artifacts during the seizures, this rule was relaxed to 8 of 10 windows.

3.2 Automatic annotation correction

An automatic approach, which is a combination of a threshold-based method and a self-confidence based method, was proposed to automatically differentiate between reliably annotated samples and incorrectly annotated samples

3.2.1 Threshold based annotation correction

Due to the large scale of the data set, the threshold-based method had been first applied to find out incorrectly annotated samples. Similar to the postprocessing, this method utilized the RMSA of every sample. The thresholds were selected based on the fact that samples whose RMSA was larger than 150 μV were very likely to be artifacts and samples whose RMSA was less than 13 μV were usually background EEG. The thresholds were then applied to the samples; the samples with an RMSA that were larger than 150 μV or were less than 13 μV were set as non-seizures by default.

3.2.2 Self-confidence based annotation correction

i. Estimation of Joint distribution

The central idea of the self-confidence-based method is to estimate the joint distribution between the observed labels and the true labels to characterize the imperfect labels, and then prune the samples that were with incorrect labels [28]. In practice, both the predicted probability of each sample and its corresponding observed label were needed to estimate the joint distribution. As the observed labels were already available, only the predicted probabilities should be estimated. In this paper, the predicted probabilities were obtained by an internal 10 folds cross-validation on the training set. Given the predicted probabilities and the observed labels, the confident joint C was first calculated. The definition of C could be written as:

$$C_{\tilde{y},y^*} = \left|\hat{X}_{\tilde{y}=i,y^*=j}\right| \tag{3}$$

where $\hat{X}_{\tilde{y}=i,y^*=j}$ is the set of samples x labelled $\tilde{y}=i$ that is likely to belong to class $y^*=j$. Determined by a pre-class threshold t_j, $\hat{X}_{\tilde{y}=i,y^*=j}$ could be written as:

$$\hat{X}_{\tilde{y}=i,y^*=j} = \left\{x \in X_{\tilde{y}=i}: \hat{p}(\tilde{y}=j;x,\theta) \geq t_j\right\} \tag{4}$$

where $j = \underset{l\in[m],\hat{p}(\tilde{y}=l;x,\theta)\geq t_l}{\arg\max} \hat{p}(\tilde{y}=l;x,\theta)$

When the predicted probability of a sample is larger than a per-class threshold, this sample would be counted as actually belonging to that threshold's class.

The threshold t_j could either be the average predicted probability of all samples in each class or a certain value decided by the knowledge of the data. In this paper, it is a certain value (0.4) that was determined by the predicted probability of a few typical incorrectly annotated samples.

With the confident joint C, the joint distribution Q could be estimated as,

$$\hat{Q}_{\tilde{y}=i,y^*=j} = \frac{\frac{C_{\tilde{y}=i,y^*=j}}{\sum_{j\in[m]} C_{\tilde{y}=i,y^*=j}}\left|X_{\tilde{y}=i}\right|}{\sum_{i\in[m],j\in[m]}\frac{C_{\tilde{y}=i,y^*=j}}{\sum_{j\in[m]} C_{\tilde{y}=i,y^*=j}}\left|X_{\tilde{y}=i}\right|} \tag{5}$$

ii. Positive-unlabelled (PU) learning

We assume that there were no seizures samples incorrectly annotated as non-seizures. This scenario could thus be considered as a PU learning problem where we only aimed to identify incorrect annotations of seizures.

Since this assumption is actually prior information of the dataset, it could be utilized in the estimation of the joint distribution Q. According to the assumption, there were no incorrectly annotated samples in the non-seizure class; hence, the $\hat{Q}_{\tilde{y}=0,y^*=1}$ should be equal to 0, and the $\hat{Q}_{\tilde{y}=0,y^*=0}$ should be equal to 1. By making the above adaption for the joint distribution, all the non-seizure samples would be considered as correctly annotated by default.

iii. Rank and removing

With the estimation of the joint distribution, the samples the training set can be ranked by their possibilities of being incorrectly annotated, e.g. the training loss or the self-confidence of their labels. In this paper, we ranked the samples with a normalized margin [29], which could be writing as.

$$\text{margin} = \hat{P}_{x,\tilde{y}=j} - \hat{P}_{x,\tilde{y}=j} \tag{6}$$

where $x \in X_{\tilde{y}=i}$

As all the samples had been ranked with their possibilities of being incorrectly annotated, the next step was to decide how many samples should be removed. In this paper, this number was estimated with the joint distribution, i.e. $N * \hat{Q}_{\tilde{y}=i,y^*=j}$ samples with largest margin $\hat{P}_{x,\tilde{y}=j} - \hat{P}_{x,\tilde{y}=i}$ should be removed from the training set. However, to keep as many training samples as possible, these samples were only removed when their predicted labels given by the internal cross validation were different from the observed labels. Finally, 1437 samples have been removed, which means about 1437 seconds seizure segments (about 1/6 of the total duration) were discarded.

3.3 Validation study and evaluation metrics

Various seizure detection approaches were compared on the fully wearable SeizeIT2 dataset. As a benchmark, the seizure detection approach, trained on full scalp EEG

Table 2. Seizure detection performance of the models trained with original full-scalp vEEG annotations, visually corrected annotations of wearable EEG data, and automatically corrected annotations of wearable EEG data

	Full scalp based labels		Visually corrected annotations		Automatically corrected annotations	
Patients	Sens [%]	FAR/h	Sens [%]	FAR/h	Sens [%]	FAR/h
1	nan	2.12	nan	0.05	nan	0.05
2	100.0	8.82	100.0	1.81	100.0	1.98
3	nan	6.62	nan	2.219	nan	1.50
4	100.0	4.56	92.3	1.24	92.3	1.16
5	100.0	1.64	100	1.00	100.0	0.55
6	nan	11.35	nan	5.51	nan	6.73
7	90.9	10.37	81.8	2.96	81.8	2.57
8	nan	3.44	nan	1.52	nan	0.76
9	100.0	1.80	100.0	0.84	100.0	0.56
10	100.0	1.72	100.0	0.22	100.0	0.48
11	100.0	8.21	100.0	3.46	100.0	2.18
12	nan	1.99	nan	0.52	nan	0.46
13	nan	0.74	nan	0.11	nan	0.11
14	100.0	14.70	100.0	4.44	100.0	3.63
15	0.0	10.10	0.0	2.41	0.0	1.25
16	100.0	9.38	100.0	0.95	100.0	0.73
17	nan	5.79	nan	1.25	nan	1.15
18	0.0	3.58	0.0	0.23	0.0	0.12
19	100.0	5.66	100.0	1.29	100.0	2.90
20	nan	2.82	nan	1.02	nan	0.98
21	100.0	0.87	100.0	0.63	100.0	0.35
22	nan	2.10	nan	0.51	nan	0.64
23	nan	4.66	nan	1.26	nan	2.07
24	nan	0.91	nan	1.24	nan	0.46
25	100.0	15.26	100.0	9.76	100.0	7.93
26	100.0	3.95	100.0	0.53	100.0	0.33
27	100.0	3.78	100.0	0.65	100.0	0.33
28	nan	8.19	nan	1.88	nan	1.00
29	100.0	2.09	100.0	0.72	100.0	0.35
30	50.0	12.82	100.0	6.41	100.0	5.59
31	100.0	17.20	100.0	8.65	100.0	6.92
32	87.5	14.33	75.0	4.05	100.0	3.18
33	50.0	7.34	50.0	2.42	50.0	1.44
34	nan	1.24	nan	0.24	nan	1.00
35	100.0	16.33	100.0	5.32	100.0	5.25
36	nan	7.23	nan	5.82	nan	2.85
37	nan	2.22	nan	0.36	nan	0.77
38	nan	10.16	nan	6.72	nan	5.05
39	nan	0.61	nan	0.02	nan	0.01
AVERAGE	85.4	6.32	86.3	2.31	87.5	1.93
MEDIAN	100.0	4.66	100.0	1.25	100.0	1.00

Abbreviations: nan, no seizures during the monitoring.

annotations, was evaluated. The training dataset (SeizeIT1) dataset contains three BTE channels, while the new test dataset (SeizeIT2) only has two BTE channels, therefore channel selection was performed to remove one unilateral channel. Then, the approach was retrained based on features derived on 2 channels. The choice of unilateral channels depended on the hemisphere where epileptic activity was most visible.

Second, the approach was retrained on SeizeIT1, but based on the visually corrected annotations from the expert. Finally, the automatic annotation correction approach, presented in this paper was applied to the SeizeIT1. The model was also retrained with the use of the automatically corrected annotations.

All results are based on patient-independent seizure detection. Two event-based metrics were used to measure the performance: seizures detection sensitivity and false alarm rate (FAR) per hour.

4. Results

4.1 Seizure detection performance

We compared the performance of the seizure detection models trained with different annotations. All the models used the same feature extraction and classification method but were trained with the 1. original full-scalp-EEG-based annotations, 2. the visual corrected BTE-EEG-based annotations, and 3. the automatically corrected annotations respectively.

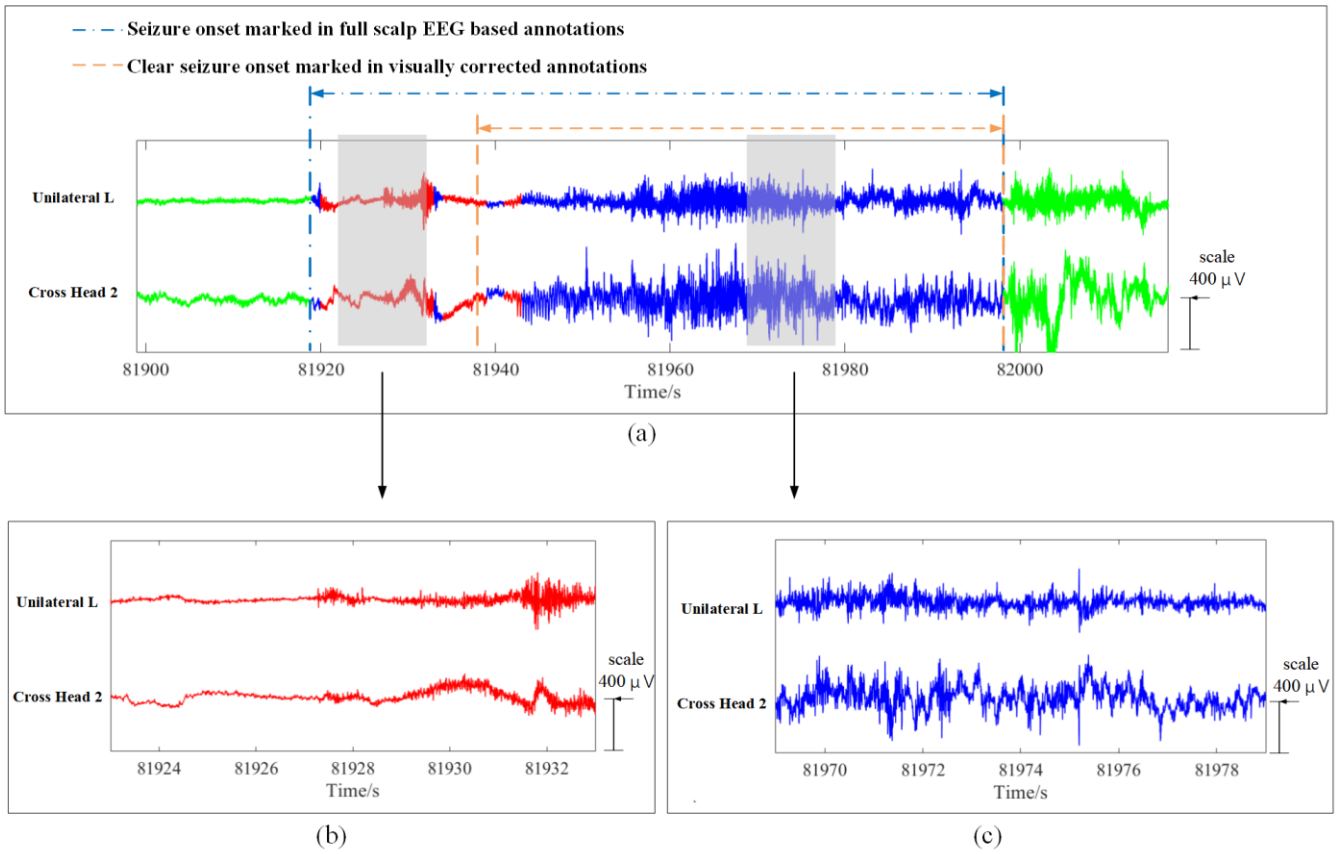


Figure 5. (a) Annotations given by different approaches on a segment of EEG signals in the BTE channel. Chain line: seizure onset marked in full scalp EEG based annotations. Dash line: seizure onset marked in visually corrected annotations. Red: seizure sub-segments marked as incorrectly annotated by automatic approach. Blue: seizure sub-segments marked as correctly annotated by automatic approach. Green: non-seizure sub-segments. (b) zoom in on incorrectly annotated sub-segments given by automatic approach. (c) zoom in on correctly annotated sub-segments given by automatic approach.

Table 2 illustrates the performances of all the 39 patients as well as their average and median. The results show that the model trained with full-scalp-EEG-based annotations could achieve 85.4% sensitivity with an average FAR of 6.32/h. The relatively high FAR is due to the fact that some artifacts were included in the seizure class of the training set during training and consequently lead to the confusion of artifacts and seizures by the algorithm. The discrimination of artifacts and seizures is challenging as the seizure patterns might be covered by artifactual patterns. This can be handled by excluding artifacts from the seizure class during training.

With the visually corrected annotations given by the epileptologist, the seizure detection model was re-trained and the results have been improved to 86.3% sensitivity and 2.31 FAR.

The model trained with the automatically corrected annotations achieves even better results (87.5% sensitivity and 1.93 FAR) compared to the model trained with the visual corrected annotations, which shows the effectiveness of the proposed automatic annotation correction approach.

Large variability of sensitivity and FAR over the patients can be observed. As shown in Table 2 the sensitivity was either 0 or very high, which suggests the high variability of seizures over different patients. For example, all the models achieved 100% sensitivity for patient 2, while all of them received 0% sensitivity for patient 15. This 0% sensitivity for patient 15 is because the only seizure of this patient was short in duration and low in amplitude which made it is difficult to be detected automatically.

4.2 Comparison and Visualization of Annotation Correction

In this subsection, we compared the original full-scalp-EEG-based annotations with the visually corrected BTE-EEG-based annotations, and the automatically corrected annotations respectively.

Specifically, the original full-scalp annotations were based on the full scalp EEG, while the visually corrected annotations were annotated by the visual analysis of the BTE-EEG. The automatic approach recognized the incorrectly annotated samples. Since every sample corresponds to a segment in EEG signals, the invisible segments during the seizures could be marked out.

In general, the neurolgoist believes that 7296 seconds of seizure segments were incorrectly annotated, while the automatic approach says there are only 1437 seconds of incorrectly annotated seizure segments. However, 90% of the automatically indificated segments are the same as those manually indificated segments; namely, the automatically indificated segments is like a subset of those manually indificated segments.

Figure 5 visualizes the annotations given by different approaches on a segment of EEG signals in the BTE channel, which includes a complete seizure and signals before and after the seizure. It can be seen that a sub-segment of EEG at the first half of the seizure was considered as incorrect annotated by the automatic approach. By zooming in on this controversial sub-segment, it can be clearly observed that this sub-segment is actually submerged in the muscle artifacts. The second half of the seizure was preserved which matches the corrected BTE-EEG-based annotations given by our expert.

It is worth noting that, some short sub-segments in the first half are also preserved, which is different from the visually corrected annotations. This is because we are conservative in the selection of hyperparameters which makes the algorithm tend to remove fewer samples. It is usually inclined to keep as many seizure samples as possible. Therefore, the approach would only mark the samples as incorrectly annotated when it is highly confident. Obviously, this also reflects the difference between how a human correct the annotations by taking into account the longer time series, and the algorithm, which assigns a probability to every epoch without taking this bigger picture into account.

5. Discussion

As manually annotating data recorded outside of the traditional hospital environments is nearly impossible, automatically monitoring epilepsy with wearable devices is becoming an increasingly important research topic. Some seizure detection approaches have been introduced specifically for wearable devices in recent years [30-34]. Most of these wearable devices used electrocardiogram (ECG) or electromyography (EMG) to detect seizures [32-34]. Despite their success in the detection of specific types of seizures (e.g. tonic-clonic seizures), they have not shown their advantages in detecting the most common focal seizures.

Compared with ECG or EMG, EEG is the gold standard of seizure detection [35]. In this paper, we aim to introduce the results of a seizure detection approach based on data recorded for the first time with a fully wearable EEG monitor and establish a framework that can handle specific challenges of training algorithms for wearable seizure detection.

First, the results on this dataset collected with the wearable Sensor Dot are similar to the results presented on the previous dataset (SeizeIT1), which simulated a BTE EEG system, as the signals from BTE positions were still collected with a traditional vEEG equipment [16]. Most importantly, the algorithm developed on the SeizeIT1 dataset was almost unaltered applied to the SeizeIT2 dataset, which guarantees full independence between training and testing of the machine learning approach. Note that the only minor difference was the presence of one extra BTE EEG derivation in SeizeIT1 dataset, which was omitted in the current version of the algorithm.

Probably the biggest challenge in developing machine learning algorithms for novel wearable devices is to obtain accurate annotation of the BTE EEG. Although simultaneous measuring BTE and vEEG could establish a ground truth on the presence of seizures, there is a high likelihood that some ictal activity is not visible on the BTE channels or that the data is corrupted by artifacts. As such, in this manuscript, we showed that training seizure detection algorithms with imperfect annotations may have a big impact on the performance of seizure detection. Even if some methods are considered to be robust to the imperfect annotations [36; 37], overfitting could not be avoided owning to their large network capacity and learning ability [19; 20]. Furthermore, the impact of different types of incorrect annotations can be different. The incorrect annotations studied in this paper are mainly caused by artifacts and seizure segments that were invisible on the wearable EEG data. Those structured deviations can have a much larger impact on classification performance than random annotating errors [38]. The classifier will learn to recognize the patterns created by those incorrectly annotated data, like when the data is correctly annotated. Our results support that an improvement in the BTE seizure detection performance can be obtained when the model is trained with more optimal annotations than the suboptimal full-scalp-based annotations.

Visual correction of the annotations can provide a much-improved data set for training the machine learning algorithm. This is because subtle ictal EEG segments and seizures with artifacts were pruned from the training set by the neurologist and therefore lead to a better training set. Specifically, removing artifacts from the seizures is advantageous to the training of the classifier, empowering it to learn the difference between an artifact and a seizure. This is in agreement with the substantial decrease in the FAR. However, it requires a well-experienced epileptologist to visually correct the annotations; other professionals might be less accurate in the correction. Moreover, re-annotating all the BTE EEG data implies a significant workload.

To establish a cost-effective framework, we further investigated the feasibility of replacing the visual work with an automatic approach. This automatic approach is a combination of a practical threshold based method and a self-confidence based method that can improve computing efficiency. The threshold based method is based on the root mean square amplitude (RMSA) and consequently is highly computationally efficient, while the self-confidence based

method is based on the predicted probability obtained by internal cross-validation and requires more computational power. Those two methods can complement each other, which means the threshold based method will process the obvious annotating error with high efficiency while the self-confidence based method will be more effective in dealing with more subtle mis-annotating.

Our results show that the proposed automatic approach achieves even better performance compared to the visual analysis carried out by the epileptologist since the proposed automatic approach focuses on finding abnormal data from a statistical perspective while the epileptologist would try to recognize established evolving ictal EEG patterns. As the proposed approach does not require any additional manual workload, the proposed approach shows its potential for further development.

While we illustrate annotation correction based on self-confidence in combination with an SVM classifier, the approach is compatible with many different classifiers and as such relevant for most other seizure detection frameworks, including dealing with other signal modalities. Future work will explore the possibility of embedding annotating correction into a deep learning architecture and combine it in our multimodal seizure detection framework [39].

6. Conclusion

This paper verified the feasibility of detecting seizures with the fully wearable device and proposes a seizure detection framework that can deal with incorrect seizure annotations. Two approaches for annotation correction, based on visual analysis and automatically respectively, have been compared on an independent dataset. Our results show the effectiveness of both methods, with a light preference for the automatic approach. As the automatic approach is much more cost-efficient and expandible to other data modalities, it is highly promising for future seizure detection systems in clinical and non-clinical environments.

7.Acknowledgements

The authors would like to thank all the patients for participating in this research. This research is funded by the EIT Health Grant: 21263 SeizeIT2 (Discreet Personalized Epileptic Seizure Detection Device). SeizeIT started out as an imec.icon research project (2016-2019), and currently continues as an EIT Health funded initiative (2020-2021). Project partners are KU Leuven, UCB Pharma, Byteflies and Pilipili, Neuroventis, Karolinska Institute, Aachen UMC, UH Freiburg, Oxford University Hospitals NHS, King's College London and CHUC Coimbra. This research is also supported by the Flemish Government: “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program and the “Deep, personalized epileptic seizure detection” FWO Research Project (G0D8321N); the Bijzonder Onderzoeksfonds KU Leuven (BOF): “Prevalentie van epilepsie en slaapstoornissen in de ziekte van Alzheimer” (C24/18/097).





8.Ethical statement

The SeizeIT1 and SeizeIT2 projects were approved by the ethical committee of UZ/KU Leuven (internal number: S59662 and S63631 respectively). A trial identifier of SeizeIT2 is also available on clinicaltrials.gov, which is NCT04284072. Written informed consent was obtained for every participant.

References

[1] French J A 2007 Refractory epilepsy: clinical overview *Epilepsia* **48** 3-7

[2] Téllez-Zenteno J F, Hernández-Ronquillo L, Buckley S, Zahagun R and Rizvi S 2014 A validation of the new definition of drug-resistant epilepsy by the I nternational L eague Against E pilepsy *Epilepsia* **55** 829-34

[3] Elger C E and Mormann F 2013 Seizure prediction and documentation--two important problems *The Lancet. Neurology* **12** 531-2

[4] Baumgartner C and Koren J P 2018 Seizure detection using scalp-EEG *Epilepsia* **59** 14-22

[5] Gu Y, Cleeren E, Dan J, Claes K, Van Paesschen W, Van Huffel S and Hunyadi B 2018 Comparison between scalp EEG and behind-the-ear EEG for development of a wearable seizure detection system for patients with focal epilepsy *Sensors* **18** 29

[6] Boeckx S, Van Paesschen W, Bonte B and Dan J. 2018. Live Demonstration: SeizeIT-A wearable multimodal epileptic seizure detection device. IEEE Biomedical Circuits and Systems Conference (BioCAS), 2018 Cleveland, Ohio, USA. IEEE,p 1.

[7] Bleichner M G and Debener S 2017 Concealed, unobtrusive ear-centered EEG acquisition: cEEGrids for transparent EEG *Front Hum Neurosci* **11** 163

[8] Jory C, Shankar R, Coker D, McLean B, Hanna J and Newman C 2016 Safe and sound? A systematic literature review of seizure detection methods for personal use *Seizure* **36** 4-15

[9] Supriya S, Siuly S, Wang H and Zhang Y 2020 Automated epilepsy detection techniques from electroencephalogram signals: a review study *Health Information Science and Systems* **8** 1-15

[10] Beniczky S, Wiebe S, Jeppesen J, Tatum W O, Brazdil M, Wang Y, Herman S T and Ryvlin P 2021 Automated seizure detection using wearable devices: A clinical practice guideline of the International League Against Epilepsy and the International Federation of Clinical Neurophysiology *Clin Neurophysiol* **132** 1173-84

[11] Swinnen L, Chatzichristos C, Jansen K, Lagae L, Depondt C, Seynaeve L, Vancaester E, Van Dycke A, Macea J, Vandecasteele K, Broux V, De Vos M and Van Paesschen W 2021 Accurate detection of typical absence seizures in adults and children using a two-channel EEG wearable behind the ears *Epilepsia*

[12] Siddiqui M K, Morales-Menendez R, Huang X and Hussain N 2020 A review of epileptic seizure detection using machine learning classifiers *Brain informatics* **7** 1-18

[13] Green R M, Messick W J, Ricotta J J, Charlton M H, Satran R, McBRIDE M M and DeWeese J A 1985 Benefits, shortcomings, and costs of EEG monitoring *Ann Surg* **201** 785

[14] Mikkelsen K B, Ebajemito J K, Bonmati-Carrion M A, Santhi N, Revell V L, Atzori G, Della Monica C, Debener S, Dijk D J and Sterr A 2019 Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy *J Sleep Res* **28** e12786

[15] Seneviratne U, Ding C, Bower S, Craig S, Leech M and Phan T G 2014 Video-based training improves the accuracy of seizure diagnosis *J Neurol Neurosurg Psychiatry* **85** 466-70

[16] Vandecasteele K, De Cooman T, Dan J, Cleeren E, Van Huffel S, Hunyadi B and Van Paesschen W 2020 Visual seizure annotation and automated seizure detection using behind-the-ear electroencephalographic channels *Epilepsia* **61** 766-75

[17] Frénay B and Verleysen M 2013 Classification in the presence of label noise: a survey *IEEE Trans. Neural Networks Learn. Syst.* **25** 845-69

[18] Garc á S, Luengo J and Herrera F 2015 *Data preprocessing in data mining* vol 72: Springer)

[19] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2021 Understanding deep learning (still) requires rethinking generalization *Commun. ACM* **64** 107-15

[20] Sukhbaatar S and Fergus R 2014 Learning from noisy labels with deep neural networks *arXiv preprint arXiv:1406.2080* **2** 4

[21] Bridge P, Fielding A, Rowntree P and Pullar A 2016 Intraobserver variability: should we worry? *J. Med. Imaging Radiat. Sci.* **47** 217-20

[22] Karimi D, Dou H, Warfield S K and Gholipour A 2020 Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis *Med Image Anal* **65** 101759

[23] EIT-Health. 2018. *SeizeIT2* [Online]. Available: https://eithealth.eu/project/seizeit2/ [Accessed 2021].

[24] Byteflies. 2021. *Personalized wearable seizure monitoring* [Online]. Available: https://www.byteflies.com/our-solutions [Accessed 2021].

[25] Seeck M, Koessler L, Bast T, Leijten F, Michel C, Baumgartner C, He B and Beniczky S 2017 The standardized EEG electrode array of the IFCN *Clin Neurophysiol* **128** 2070-7

[26] Hunyadi B, Signoretto M, Van Paesschen W, Suykens J, Van Huffel S and De Vos M 2012 Incorporating structural information from the multichannel EEG improves patient-specific seizure detection *Clin Neurophysiol* **123**

[27] Greene B R, Faul S, Marnane W P, Lightbody G, Korotchikova I and Boylan G B 2008 A comparison of quantitative EEG features for neonatal seizure detection *Clin Neurophysiol* **119** 1248-61

[28] Northcutt C, Jiang L and Chuang I 2021 Confident learning: Estimating uncertainty in dataset labels *J Artif Intell Res* **70** 1373-411

[29] Wei C, Lee J, Liu Q and Ma T 2018 On the margin theory of feedforward neural networks *arXiv preprint arXiv:1810.05369*

[30] Van Paesschen W 2018 The future of seizure detection *The Lancet Neurology* **17** 200-2

[31] Becker T, Vandecasteele K, Chatzichristos C, Van Paesschen W, Valkenborg D, Van Huffel S and De Vos M 2021 Classification with a deferral option and low-trust filtering for automated seizure detection *Sensors* **21** 1046

[32] Miloevi M, Vel A, Bonroy B, Ceulemans B, Lagae L, Vanrumste B and Huffel S V 2016 Automated Detection of Tonic–Clonic Seizures Using 3-D Accelerometry and Surface Electromyography in Pediatric Patients *IEEE. J Biomed Health* **20** 1333-41

[33] Kusmakar S, Karmakar C K, Yan B, O'Brien T J, Muthuganapathy R and Palaniswami M 2018 Automated detection of convulsive seizures using a wearable accelerometer device *IEEE Trans Biomed Eng* **66** 421-32

[34] Beniczky S, Arbune A A, Jeppesen J and Ryvlin P 2020 Biomarkers of seizure severity derived from wearable devices *Epilepsia* **61** S61-S6

[35] Sharmila A 2018 Epilepsy detection from EEG signals: a review *J Med Eng Technol* **42** 368-80

[36] Rolnick D, Veit A, Belongie S and Shavit N 2017 Deep learning is robust to massive label noise *arXiv preprint arXiv:1705.10694*

[37] Drory A, Ratzon O, Avidan S and Giryes R 2018 The Resistance to Label Noise in K-NN and DNN Depends on its Concentration *arXiv preprint arXiv:1803.11410*

[38] Algan G and Ulusoy L 2020 Label Noise Types and Their Effects on Deep Learning *arXiv preprint arXiv:2003.10471*

[39] Vandecasteele K, De Cooman T, Chatzichristos C, Cleeren E, Swinnen L, Macea Ortiz J, Van Huffel S, Dümpelmann M, Schulze-Bonhage A and De Vos M 2021 The power of ECG in multimodal patient-specific seizure monitoring: Added value to an EEG-based detector using limited channels *Epilepsia*