 doi:10.20944/preprints202012.0527.v1

Article

Classification with a Deferral Option and Low-Trust Filtering for Automated Seizure Detection

Thijs Becker [1,*], Kaat Vandecasteele [2], Christos Chatzichristos [2], Wim Van Paesschen [3,4], Dirk Valkenborg [1], Sabine Van Huffel [2] and Maarten De Vos [2,5]

1 I-Biostat, Data Science Institute, Hasselt University, Hasselt 3500, Belgium; dirk.valkenborg@uhasselt.be (D.V.)
2 Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven 3001, Belgium; kaat.vandecasteele@esat.kuleuven.be (K.V.); christos.chatzichristos@esat.kuleuven.be (C.C.); sabine.vanhuffel@esat.kuleuven.be (S.V.H.); maarten.devos@kuleuven.be (M.D.V.)
3 Department of Neurology, UZ Leuven, Leuven 3001, Belgium; wim.vanpaesschen@uzleuven.be (W.V.P.)
4 Laboratory of Epilepsy Research, KU Leuven, Leuven 3001, Belgium
5 Department of Development and Regeneration, KU Leuven, Leuven 3001, Belgium
* Correspondence: thijs.becker@uhasselt.be

Abstract: Wearable technology will become available and allow prolonged electroencephalography (EEG) monitoring in the home environment of patients with epilepsy. Neurologists analyse the EEG visually and annotate all seizures, which patients often under report. Visual analysis of a 24 hour EEG recording typically takes one to two hours. Reliable automated seizure detection algorithms will be crucial to reduce this analysis. We study a dataset of behind-the-ear EEG measurements. Our first aim was to develop a methodology to reduce the EEG dataset by classifying part of the data automatically, while retaining 100% detection sensitivity (DS). Prediction confidences are determined by temperature scaling of the classification model outputs and trust scores. A DS of approximately 90% (99%) can be achieved when automatically classifying around 90% (60%) of the data. Perfect DS can be achieved when automatically classifying 50% of the data. Our second contribution demonstrates that a common modelling strategy, where predictions from several short EEG segments are used to obtain a final prediction, can be improved by filtering out untrustworthy segments with low trust scores. The false detection rate shows a relative decrease between 21% and 43%, and the DS shows a small increase or decrease.

Keywords: epilepsy; seizure detection; electroencephalography; classification with a deferral option; home monitoring; long-term monitoring; wearables

1. Introduction

Epilepsy is a neurological disorder that affects around 0.8% of the population worldwide [1]. Epilepsy patients have recurrent unprovoked seizures, which significantly affect their quality of life. Anti-epileptic drugs provide adequate treatment for 70% of the patients [2]. The seizure burden is an important variable for treatment decisions and the evaluation of drug trials. One should therefore be able to objectively document seizures occurring over a time span of days to weeks [3], preferably in a home environment [4]. Unfortunately, seizure reporting by patients is unreliable [5]. Seizures are therefore detected with devices that record biosignals, most commonly full scalp electroencephalography (EEG) [6], which is uncomfortable to wear for a long period of time. Other biosignals such as electrocardiograms, electromyograms, accelerometry, and EEG from behind-the-ear sensors can be used outside the hospital [7]. They have the advantage that those measuring devices are more tolerated when being worn for an extended period of time. Combining several of these biosignals can improve seizure detection performance [8,9]. We are currently performing the SeizeIT2 study,

 doi:10.20944/preprints202012.0527.v1

which is a multicenter study to examine clinical scenarios for long-term monitoring of epileptic seizures with a wearable biopotential technology in the home environment [10] (ClinicalTrials.gov Identifier: NCT04284072).

Regardless of the measured biosignal(s), manual analysis of the output is a time-consuming task. Automated seizure detection with machine learning has therefore received a lot of attention [11–13]. Impressive results have been obtained, though mostly on retrospective single-center datasets, and only for certain seizures types. For focal seizures, no results were achieved that are good enough to be applied in practice.

Achieving a performance of automated seizure detection that is sufficient for implementation in a clinical setting remains challenging. Automated seizure detectors that are commercially available are reported to have a low detection sensitivity (DS): using full scalp EEG they detected at least one seizure in only 53% of measurements containing seizures [14]. Recently, a seizure detection competition was held on the Temple University Hospital Seizure Detection Corpus, which is the largest open source corpus of its type and includes representative cases of different types of seizures [15]. Despite the size of the dataset and the use of advanced machine learning algorithms, the participants were unable to achieve a performance that is sufficient for clinical practice and that could be used for all types of seizures [16].

We investigate a dataset containing EEG measurement from four behind-the-ear sensors, from our SeizeIT1 study [17,18]. Behind-the-ear sensors are able to detect epileptic seizures for focal onset and generalized seizure types [19,20]. In contrast to full scalp EEG, there are only a few studies that investigate automated seizure detection algorithms on behind-the-ear EEG [20–22].

Our first contribution in this paper, is the evaluation of the performance of the classifier in case that the EEG segments for which the classifier is least confident are deferred to a human annotator, who is assumed to annotate perfectly. A similar scenario is quite common in clinical epilepsy research: the algorithm flags all suspicious activity, which is then presented to a human annotator [23]. Learning algorithms with a reject option have a long history in machine learning research [24,25]. Classification with the option to defer to a human expert is receiving increasing attention in the current AI literature [26–29], and is particularly relevant for medical tasks [30–32]. Application of this approach to seizure detection has been limited. Computer-assisted detection of epileptic discharges from full scalp EEG has been investigated by Clarke et al. [23]. On a retrospective dataset, a neural network achieved a DS of 96.7% with a false detection rate (FDR) per 24 hours of 1670. They employed this model in a clinical application of ambulatory measurement of 7 patients with idiopathic generalized epilepsy. 10-second EEG segments that contained a seizure detection were deferred to a human annotator. The data that needed to be reviewed was reduced to between 60% and 90% of the full data. The precision and FDR were tuned by changing the classification threshold of the network. The DS was not measured for the clinical application. In this article, we defer EEG segments with a length of at least 5 minutes to a human annotator. We investigate the support vector machine (SVM) models that were trained to perform automated seizure detection on patients with focal epilepsy in [21]. Two patient-independent models are considered, with detection sensitivities of 64.1% and 83.0%, and false detection rates (FDR) per 24 hours of 2.9 and 17.2, respectively. We use the confidences derived from the SVM output or from so-called trust scores [28] (which only depend on the labels and on the features derived from the EEG to train the SVM). Segments that contain a seizure detection are always deferred, so the FDR is always 0. For both SVMs, we achieve a DS of approximately 90% (99%) after deferring around 10% (40%) of the data. Perfect performance can be achieved when deferring 50% of the data. These results indicate that it is of interest to investigate algorithms that combine a good classification performance with good confidence estimates, instead of focusing solely on the performance when all data is classified by the algorithm.

 doi:10.20944/preprints202012.0527.v1

Our second contribution demonstrates that a common modelling strategy, where predictions from several short segments are used to obtain a final prediction [9,20,21,33], can be improved by filtering out untrustworthy segments. The FDR shows a relative decrease between 21% and 43%, and the positive predictive value and F1-score improve considerably. The detection sensitivity shows a small increase or decrease. This filtering approach only works with the trust scores. For the SVM confidences it lowers the performance.

2. Materials and Methods

Most code was implemented in a Conda environment in Python 3. The main libraries that were used are scikit-learn [34] and SciPy [35]. Our code is made available at https://github.com/thijsrmbecker/classify_w_deferral_seizure. Some preliminary data processing was done in Matlab.

2.1. Dataset

The dataset consists of recordings from a traditional 10-20 scalp EEG with four extra behind-the-ear electrodes [18,20,36], as shown in Figure 1. The neurologist (W.V.P.) annotated all seizures on the standard video-EEG recordings. In the present study, we only used the behind-the-ear EEG measurements plus the seizure annotations of the standard video-EEG recordings, and call these "full seizure events". We only took into consideration the seizure segments that could be blindly annotated by the neurologist (W.V.P.) in the behind-the-ear measurements, which are 63% of all seizures. These included mainly patients with focal seizures from temporal lobe origin or patients with other focal epilepsy syndromes with ictal propagation that was picked up by the behind-the-ear electrodes, and two patients with focal to bilateral tonic clonic seizures. 54 patients are included, 42 of which had seizures during the measurements. The dataset consists of approximately 220 days of EEG time series and contains 114 seizures. We refer to [21] for a detailed discussion of the content of the dataset.

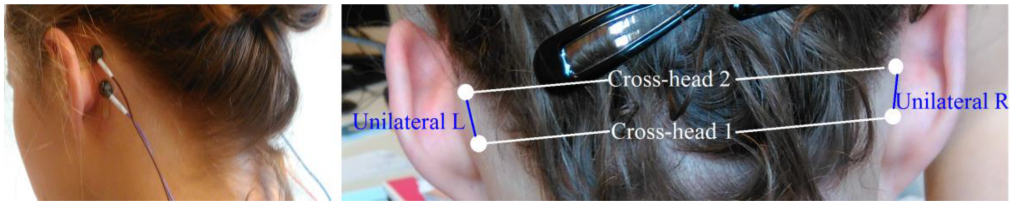

Figure 1. Behind-the-ear electroencephalographic setup. Left panel shows extra behind-the-ear electrodes glued to the skin. Right panel shows bipolar channel derivations. Reproduced with permission from Gu et al. [20].

2.2. Models

We investigate patient-independent SVMs as presented in [21]. The goal of the models is to detect focal seizures visible on the behind-the-ear EEG by the neurologists, and with a length of at least 10 seconds. Features are extracted from 2-second segments with 50% overlap. Seizures segments have label 1 and seizure-free segments have label 0. If more than 7 out of 10 subsequent 2-second segments are classified as a seizure by the SVM, the 10-second segment is classified as a seizure (also referred to as a seizure flag). If there are subsequent seizure flags, only the first flag is retained.

For each seizure a 10-second EEG segment that contains a clear ictal EEG pattern was selected. The SVM model from [21] was trained on these 10-second segments with a clear ictal EEG pattern. This procedure ensured that each seizure has the same importance when training the SVM, independent of its total length. An example of the full seizure labels and the clear ictal labels is shown in Figure 2. The SVM trained on the clear ictal segments has a FDR per 24 hours of 2.9 and a DS of 64.1%. It is referred to as clear ictal SVM (CI SVM). We also investigate an SVM model trained on the full seizure

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

labels. It has a FDR per 24 hours of 17.2 and a DS of 83.0%. It is referred to as full seizure SVM (FS SVM).

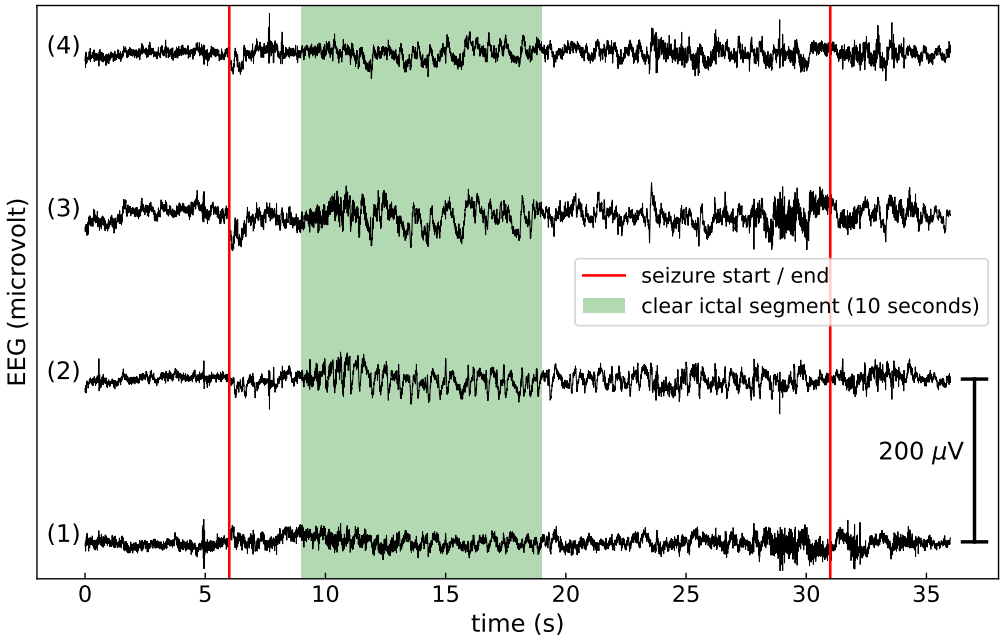

Figure 2. An example of the full seizure and clear ictal labels, used to train the full seizure SVM and clear ictal SVM. (1) crosshead 1, (2) crosshead 2, (3) unilateral left, (4) unilateral right. The full seizure as annotated by the neurologists occurs between the vertical red lines. A 10-second EEG segment that contains a clear ictal pattern is also annotated.

2.3. Confidence Measures

The confidence in the prediction of the SVMs is derived from the distance d to their separating hyper-plane. Temperature scaling [37] is performed to convert these distances to probabilities:

$$p_a = \frac{1}{1 + \exp(-a.d)}, \tag{1}$$

where $a > 0$ is a parameter which can be optimized. Temperature scaling is a simplified version of Platt scaling [38], which is often used to obtain probabilities from SVM outputs. In contrast to Platt scaling, the class of the predictions cannot change with temperature scaling. Temperature scaling is surprisingly effective at calibrating neural networks [37]. Because the classification threshold is at $p_a = 0.5$ (i.e., $d = 0$), the confidence is calculated as $|p_a - 0.5|$.

Trust scores are introduced in [28]. They are calculated with the code from https://github.com/google/TrustScore. The algorithm consists of 2 steps. In the first step one calculates the α-high-density set of each class. This is done by removing the α-fraction of the samples with the lowest density of each class (which may be outliers). The samples in the α-high-density set can be interpreted as the "representative" fraction of each class. α is a variable determined by the user, which can be optimized. In the second step, the trust score of each point in the test set is found by calculating its (Euclidean) distance to the closest point of the α-high-density set of the other class, and dividing it by the distance to the closest point of the α-high-density set of the predicted class.

We skip the first step when calculating the trust scores. This has two important advantages: It makes the calculation significantly faster; and it eliminates the only 2 hyper-parameters associated with the trust model (α and a parameter k to estimate the empirical density based on k-nearest neighbours). The disadvantage is that we could lose some performance. The original 67 features extracted from the EEG time series [21]

 doi:10.20944/preprints202012.0527.v1

are reduced to 20 dimensions using principal component analysis. We did not optimize for the number of principal components. We took this number because it was also used in the original article on trust scores [28] .

A low trust score can be interpreted as a sample that is atypical for its predicted class. This could be a result of the presence of noise or artifacts in the EEG, but could also be caused by other reasons. Under some distributional assumptions one can show that a high (low) trust score implies that the classifier likely agrees (disagrees) with the Bayes-optimal classifier [28]. Because trust scores are independent of the classifier, they can be used in conjunction with any classifier. We train the trust models using either the full seizure or clear ictal labels.

Trust models are fitted on a subset of the data. All seizure segments are included. For each patient, one-minute long non-seizure segments are selected 15 minutes apart, with each segment containing 30 non-overlapping 2-second segments. Out of all these selected one-minute long segments, we randomly draw 100 for each patient. If there is not enough data to randomly draw 100 segments, the one-minute segments are selected 5 minutes apart. This procedure is done to assure that fitting the trust models and calculating the trust scores is fast enough.

2.4. Performance Metrics

If a seizure flag occurs between the onset and the end of a seizure, it counts as a true positive (TP). If no seizure flags occur between the onset and the end of a seizure it is a false negative (FN). All seizure flags that do not overlap with a seizure are false positives (FP). The performance metrics are calculated as in [21]: detection sensitivity = TP / (TP + FN); false detection rate = FP / recording length, where FP within 10 seconds of each other are counted as one FP; positive predictive value (PPV) = TP / (TP + FP) (also called precision); and F1-score = 2 TP / (2 TP + FP + FN). The PPV and F1-score are only calculated on patients that have seizures. The detection delay is the time difference in seconds between the start of the seizure and the seizure flag. All performance metrics are calculated per patient and then averaged.

2.5. Classification With A Deferral Option

If a segment is annotated by a human, we assume it is done perfectly. The full EEG signal is divided into segments that can be deferred to a human annotator as follows. All 10-second seizure flags are put in the middle of a 5-minute segment. If there is less than 5 minutes between two such 5-minute segments, they are merged. Afterwards, the remaining EEG signal (that does not contain seizure flags) is divided into 5-minute segments. Segments that contain a seizure flag are always deferred. Other segments are deferred based on their confidence score. The deferral strategy is visualised in Figure 3.

The confidence score of each segment is calculated as follows. We consider the trust scores or SVM confidences (referred to as score in this Section) of all 2-second segments in the segment. We either calculate the average of the scores of all 2-second segments, or we take the percentage p_{low} of 2-second segments with the lowest score, and calculate the average of only those 2-second segments. p_{low} is a hyper-parameter that can be optimized.

The same percentile of lowest-confident segments are deferred for each patient. After deferring these segments, remaining adjacent segments are merged. If a seizure is in a deferred segment for at least 10 seconds, it counts as detected; the classification of the part (if any) that is annotated by the algorithm does not influence the performance. If an undetected seizure is split in such a way that both the deferred and non-deferred part contain less than 10 seconds, it counts as undetected.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

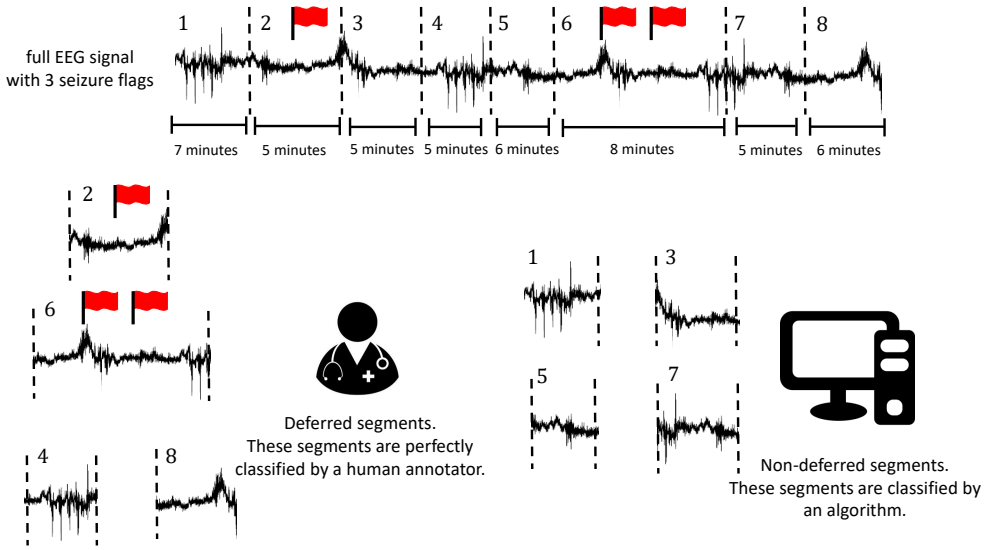

Figure 3. Visualisation of the deferral scheme. EEG segments are deferred to a human annotator or they are classified by an algorithm. The human annotator is assumed to annotate perfectly. Segments that contain seizure flags are always deferred. The minimum length of a segment is 5 minutes, but they can be longer (e.g. if they contain several seizure flags or are on the boundary). For the example in this figure, segments 2 and 6 are automatically deferred because they contain seizure flags. Segments 4 and 8 are deferred because the classifier is the least confident regarding its classification for these segments. Segments 1, 3, 5, and 7 are classified by the algorithm.

2.6. Low-Trust Filtering

From now on, we refer to 2-second segments simply as segments while all the other segment lengths will always be explicitly specified. In the original algorithm [21], a 10-second EEG segment is classified as a seizure if more than 7 out of 10 of the segments are classified as a seizure. This classification is influenced by segments that are difficult to classify (noise, artifacts, ...) and could, potentially, negatively influence the performance. Since we have 10 predictions, we can remove these "untrustworthy" predictions while still classifying all 10-second segments. We call this approach low-trust filtering (LTF). The new classification rule is as follows: If less than 5 out of 10 predictions are removed, a seizure flag is created if the mean of the remaining predictions is greater than 0.7; otherwise, a seizure flag is created if the mean of the 5 highest-trusted segments is greater than 0.7. This classification rule can be seen as a generalisation of the rule from [21], with the requirement that at least 5 predictions should be taken into account. We did not optimize for this classification rule, but tried out two variations of this rule in early experiments. These other two classification rules, and the arguments for our final choice, are discussed in the supplementary material.

Cross-validation is performed with the F1-score. Both parameters (a and the percentile of segments to filter) are selected through a grid search approach. The same percentile of segments is filtered for each patient.

For LTF with trust scores, the only parameter that needs to defined is the optimal percentile of lowest-trust segments to filter. This has to be done on a validation set containing patients which were not used to train the trust model. To avoid over-fitting, we perform nested cross-validation [39]. This is a conservative approach to cross-validation [40], so we do not expect any possibility of over-fitting. The 54 patients are divided into 6 folds containing 9 patients each. The total dataset contains approximately 220 days of EEG time series, with a total of 490 10-second seizure segments (obtained from 114 seizure events). Folds are randomly created, with the constraints that each fold should contain between 65 and 96 10-second seizure segments, and between 23.1 and 48.6 days of EEG data. The cross-validation approach is visualised in Figure 4. Each fold

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

is used as the test set once. It is the performance on this set that is reported. We explain the procedure when the test set contains the patients of the first fold. We merge folds 2 and 3 ($fold_{23}$), and folds 4 and 5 ($fold_{45}$). Fold 6 is split in half and each half is merged with $fold_{23}$ or $fold_{45}$. The end result is one fold containing 22 patients and another fold containing 23 patients, with a good balance of number of seizures and amount of EEG data between the folds. We perform 2-fold cross-validation with $fold_{23}$ and $fold_{45}$: we fit a trust model on the patients in $fold_{23}$ and $fold_{45}$, and determine the optimal percentile to filter on $fold_{45}$ and $fold_{23}$, respectively. The average of these two optimal percentiles to filter is used to determine the performance on the test fold.

After using cross-validation to determine the optimal percentile of lowest-trust segments to filter, we still have to determine which segments to filter. To achieve optimal performance, this is done with a trust model fitted on all 53 patients besides the patient under consideration.

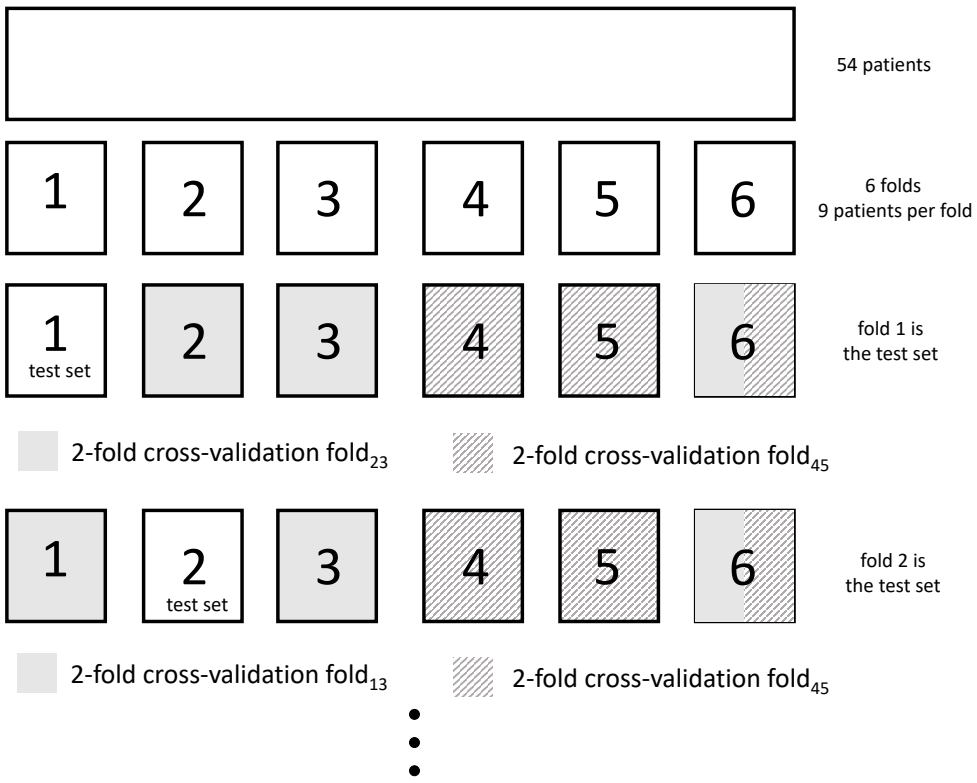

Figure 4. Visualisation of the cross-validation approach for low-trust filtering with trust scores. The 54 patients are divided into 6 folds. Each fold is used as the test set once. The optimal percentile of lowest-trust segments to remove is determined with 2-fold cross-validation. The determination of the folds for cross-validation is visualised when fold 1 or fold 2 are the test set.

For LTF with the SVM confidence scores we use leave-one-patient-out cross-validation: the optimal parameters are determined on all patients except the patient under consideration. We optimize on both the lowest-trust percentile to filter and the a variable from equation (1). Taking the distances to the separating hyper-plane d without performing temperature scaling is also considered.

We test the null hypothesis that the distribution of the performances is the same with the one-sided Wilcoxon signed-rank test (one-sided because we test whether the performance is greater or smaller). This is a paired difference test that compares the change in performance before and after LTF for each patient. We reject the null hypothesis if the p-value is lower than 0.05.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

3. Results

The main results are discussed here. Additional results are presented in the supplementary material.

3.1. Classification With A Deferral Option

For the SVM confidences we use the distances to the separating hyper-plane. These provide the same results as temperature scaling.

An example of the behaviour of the DS for different p_{low} is shown in Figure 5. The optimal p_{low} lies between 1 and 10. $p_{\text{low}} = 100$ performs the worst. From now on, we report results for $p_{\text{low}} = 5$.

The results for the CI SVM are shown in Figure 6. After deferring all segments that contain a seizure flag, which is approximately 1% of the data, we get a FDR of 0. A DS of 89% (99%) is reached after deferring 11% (36%) of the data. Perfect performance on all considered metrics is reached when deferring 50% of the data. SVM confidences slightly outperform trust scores.

The results for the FS SVM are shown in Figure 7. After deferring all segments that contain a seizure flag, which is approximately 4.5% of the data, we get a FDR of 0. A DS of 90% (99%) is reached after deferring 9% (38%) of the data. In this case, trust scores slightly outperform SVM confidences. These results are very similar to the CI SVM, despite the difference in performance when no data is deferred. Perfect performance on all considered metrics is reached when deferring 62% of the data. This is worse compared to the CI SVM. However, the improvement from 99% to 100% DS is determined by the detection of one seizure. When this is achieved is expected to be subject to quite some random variation. See, e.g., the variation in DS for different p_{low} in Figure 5.

The number and average length of the deferred segments as a function of the fraction of deferred data is plotted for the CI SVM in Figure 8. This is normalized per patient and per 24 hours of EEG data. The average length of a deferred segment starts at around 5 minutes for deferral percentages close to 0, as expected. Because adjacent deferred segments are merged, the average segment length increases for higher deferral percentages. At 10% deferral around 20 segments with an average length of 10 minutes are deferred. At 40% deferral we defer around 40 segments with an average length of 15 minutes. The maximum number of deferred segments is approximately 45 and is reached at around 50% deferral. The behaviour of the FS SVM is similar, as shown in the supplementary material.

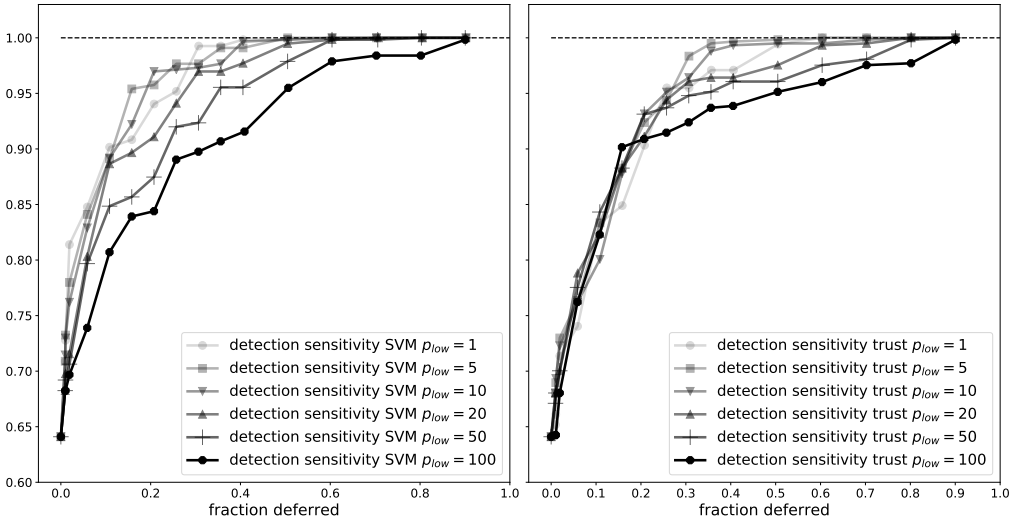

Figure 5. The average detection sensitivity as a function of the fraction of the data that is deferred to a human annotator, for the CI SVM, for different p_{low}. Segments are deferred using SVM confidences (left) and trust scores from a trust model train on the FS labels (right).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

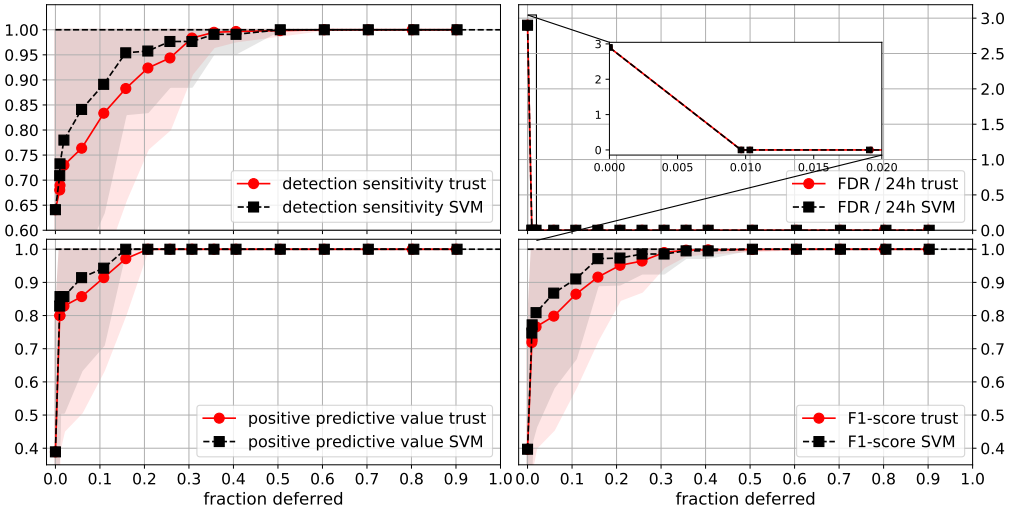

Figure 6. Average performance as a function of the fraction of the data that is deferred to a human annotator, for the CI SVM ($p_{\text{low}} = 5$). The standard deviation of the performance is shown as a shaded area, with the upper values capped at one. Segments are deferred using the SVM confidences (SVM) or trust scores (trust) from a trust model trained on the FS labels. The first point with fraction deferred larger than zero is the performance when all segments that contain a seizure flag are deferred. The inset plotted in the FDR figure shows that around 1% of the EEG data is contained in segments that contain seizure flags.

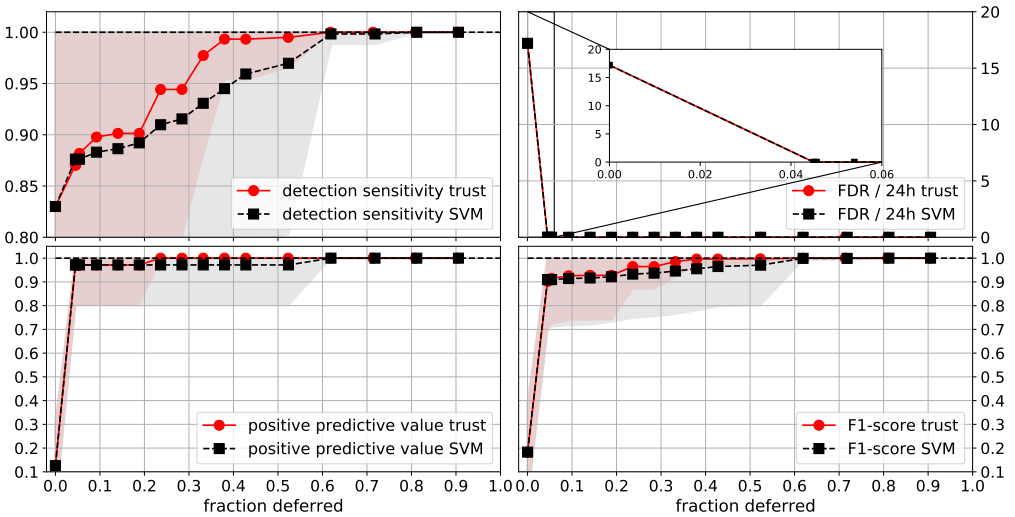

Figure 7. Average performance as a function of the fraction of the data that is deferred to a human annotator, for the FS SVM ($p_{\text{low}} = 5$). The standard deviation of the performance is shown as a shaded area, with the upper values capped at one. Segments are deferred using the SVM confidences (SVM) or trust scores (trust) from a trust model trained on the FS labels. The first point with fraction deferred larger than zero is the performance when all segments that contain a seizure flag are deferred. The inset plotted in the FDR figure shows that around 4.5% of the EEG data is contained in segments that contain seizure flags.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

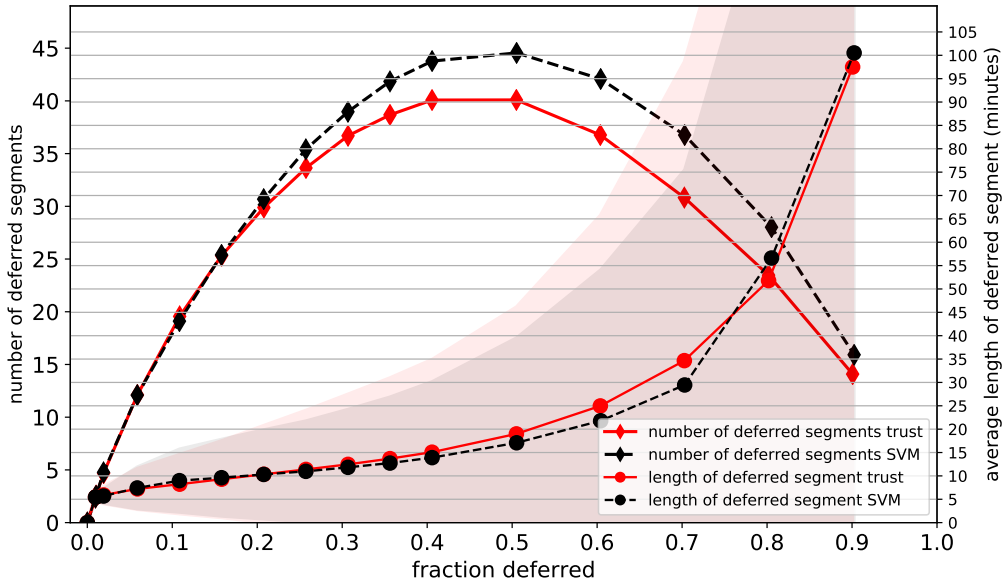

Figure 8. Average number and average length (minutes) of the deferred segments, as a function of the fraction of the data that is deferred to a human annotator, for the CI SVM ($p_{\text{low}} = 5$). The standard deviation of the length is shown as a shaded area, with the lower values capped at zero. Segments are deferred using the SVM confidences (SVM) and trust scores (trust) from a trust model trained on the FS labels.

3.2. Low-Trust Filtering

We first discuss the results for the CI SVM, see Table 1 for the average performance. If one performs LTF with the trust models fitted on the CI labels, the DS decreases slightly but non-significantly. The FDR per 24 hours decreases from 2.9 to 1.7, a relative decrease of 41%. The PPV goes from 38.9% to 48.6%, and the F1-score from 39.7% to 49.2%. If the FS labels are used to fit the trust models, LTF lowers the FDR and increases the DS. Both improvements are significant. The PPV is 50.8% and the F1-score is 52.1%. LTF with the CI SVM confidences does not improve the performance, and in fact degrades the FDR. The distances d and temperature scaling optimized for a provide the same results. On average, 2% of the lowest-trust segments are filtered for LTF with the trust models trained on the FS labels. Similar behaviour is observed for the median performances, see Table 2.

Table 1: Mean (standard deviation) performance of: CI SVM, LTF with the confidences of this SVM (CI SVM conf.), and LTF with trust models trained on CI (trust CI) and FS (trust FS) labels. A statistically significant improvement or degradation in performance compared to the original SVM is denoted by a star. The best result (or not significantly different from the best result) is shown in bold.

metric \ method	CI SVM	LTF, trust CI	LTF, CI SVM conf.	LTF, trust FS
DS (%)	64.1 (41.5)	63.8 (41.0)	64.1 (41.5)	**71.4*** (**38.6**)
FDR / 24 h	2.9 (5.6)	**1.7*** (**3.8**)	5.4* (11.7)	2.3* (4.7)
PPV (%)	38.9 (38.9)	**48.6*** (**40, 5**)	38.7* (39.0)	**50.8*** (**40.6**)
F1-score (%)	39.7 (34.2)	**49.2*** (**36.9**)	39.4* (34.3)	**52.1*** (**36.5**)
detection delay (s)	22.1 (13.2)	23.2* (12.2)	**21.9*** (**13.0**)	**21.4** (**12.0**)

 doi:10.20944/preprints202012.0527.v1

Table 2: Median [range] of performance of: CI SVM, LTF with the confidences of this SVM (CI SVM conf.), and LTF with trust models trained on CI (trust CI) and FS (trust FS) labels.

metric \ method	CI SVM	LTF, trust CI	LTF, CI SVM conf.	LTF, trust FS
DS (%)	100 [0, 100]	83.3 [0, 100]	100 [0, 100]	100 [0, 100]
FDR / 24 h	1.2 [0, 31.5]	0.46 [0, 20.5]	2.0 [0, 66.5]	0.58 [0, 24.3]
PPV (%)	23.1 [0, 100]	40.0 [0, 100]	23,1 [0, 100]	37.5 [0, 100]
F1-score (%)	31.6 [0, 100]	50.0 [0, 100]	31.6 [0, 100]	50.0 [0, 100]
detection delay (s)	19.3 [2, 55]	20.3 [3, 56]	19.3 [2, 55]	18.3 [3, 56]

The results for the FS SVM are given in Table 3 for the average performance. For the trust models trained on the FS labels, the FDR per 24 hours decreases from 17.2 to 10.6, a relative decrease of 38%. The DS decreases slightly but non-significantly. The PPV goes from 12.6% to 20.3% and the F1-score goes from 18.3% to 27.7%. LTF with the trust models trained on the CI labels performs slightly better. LTF with the FS SVM confidences does not improve the performance, and in fact degrades the FDR. The distances d and temperature scaling optimized for a give the same results. On average, 10% of the lowest-trust segments are filtered for LTF with the trust models trained on CI labels. Similar behaviour is observed for the median performance, see Table 4.

Table 3: Mean (standard deviation) performance of: FS SVM, LTF with the confidences of this SVM (FS SVM conf.), and LTF with trust models trained on CI (trust CI) and FS (trust FS) labels. A statistically significant improvement or degradation in performance compared to the original SVM is denoted by a star. The best result (or not significantly different from the best result) is shown in bold.

metric \ method	FS SVM	LTF, trust FS	LTF, FS SVM conf.	LTF, trust CI
DS (%)	**83.0** (**30.0**)	**81.9** (**10.6**)	**83.0** (**30.0**)	**81.6** (**32.3**)
FDR / 24 h	17.2 (21.0)	10.6* (14.6)	31.6* (42.4)	**9.8*** (**14.2**)
PPV (%)	12.6 (17.4)	20.3* (23.9)	12.6 (17.4)	**23.8*** (**28.3**)
F1-score (%)	18.3 (20.3)	27.7* (27.1)	18.3 (20.3)	**30.2*** (**29.6**)
detection delay (s)	**21.6** (**17.4**)	22.2* (18.0)	**21.6** (**17.4**)	22.8* (18.0)

Table 4: Median [range] of performance of: FS SVM, LTF with the confidences of this SVM (FS SVM conf.), and LTF with trust models trained on CI (trust CI) and FS (trust FS) labels.

metric \ method	FS SVM	LTF, trust FS	LTF, FS SVM conf.	LTF, trust CI
DS (%)	100 [0, 100]	100 [0, 100]	100 [0, 100]	100 [0, 100]
FDR / 24 h	9.54 [0.26, 98.2]	5.3 [0, 75.8]	16.3 [0.4, 205.7]	4.5 [0, 70.3]
PPV (%)	5.6 [0, 87.5]	12.5 [0, 100]	5.6 [0, 87.5]	12.5 [0, 100]
F1-score (%)	10.5 [0, 87.5]	16.7 [0, 100]	10.5 [0, 87.5]	20.3 [0, 100]
detection delay (s)	16.9 [1, 89]	16.4 [2, 90]	16.9 [1, 89]	16.6 [2, 90]

The LTF results are similar for both SVM models. LTF with trust models leads to a relative decrease of the FDR of around 40% for 3 out of the 4 cases. For those cases a small decrease of the DS is observed. The other case (CI SVM with a trust model trained on the FS labels) the FDR shows a relative decrease of approximately 20%, with a small increase of the DS. The strong decrease in FDR leads to strong improvements in the PPV and F1-score. No pronounced effects are observed for the detection delays. In both cases, LTF with the trust models trained on the labels different from the labels used to train the SVM slightly outperforms the other trust model. SVM confidences do not improve the performance, and in fact increase the FDR.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

Figure 9 shows a visualisation of a false positive segment that is removed after performing LTF. In this seizure-free EEG signal, there is a part where the SVM classifies many segments as a seizure. All but one of these seizure predictions are deemed untrustworthy by the trust scores, so the seizure flag disappears after LTF.

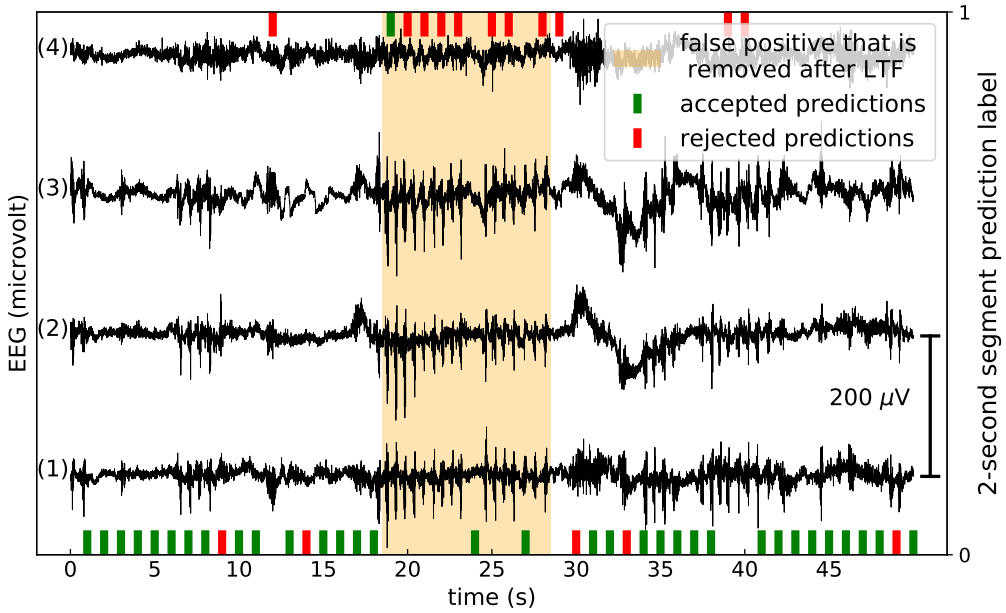

Figure 9. A visualisation of a false positive that is removed after low-trust filtering, for the CI SVM and a trust model trained on the FS labels. (1) crosshead 1, (2) crosshead 2, (3) unilateral left, (4) unilateral right. If the SVM predicts 0 a bar on the x-axis in the middle of the corresponding 2-second segment is shown. If the SVM predicts 1 a bar on top of the figure is shown. Predictions that are flagged as untrustworthy are shown as red bars, otherwise they are green. The seizure flag that disappears after low-trust filtering is shown in orange.

3.3. Difference Between SVM Confidences and Trust Scores

We compare the rankings of the SVM confidences and the trust scores with the Kendall rank correlation coefficient (KRCC). The KRCC is calculated between the SVM confidences and trust scores for each patient and then averaged. For the CI SVM confidences the KRCC is 0.16 with the trust model trained on the CI labels, and 0.17 with trust model trained on the FS labels. For the FS SVM confidences the KRCC is 0.16 with the trust model trained on the FS labels, and 0.11 with the trust model trained on the CI labels. These KRCC values are low, and therefore show that the rankings of the SVM confidences and trust scores differ notably.

Trust scores are superior for removing false positives for both LTF and classification with a deferral option. This can be seen in Tables 1, 2, 3, and 4 for LTF. For classification with a deferral option, this is shown in Figure 10. In this figure we plot the FDR as a function of the fraction of the data that is deferred to a human annotator. In contrast to our main approach, the segments that contain seizure flags are not automatically the first to be deferred. If a seizure flag is in a deferred segment for at least one second, we assume that it is completely checked by the human annotator, even if a part is checked by the algorithm. We clearly see that trust scores are better at detecting segments that contain false positives. For detecting 5-minute segments that contain false negatives our results suggest that SVM confidences and trust scores behave similarly. Either SVM confidences (CI SVM, Figure 6) or trust scores (FS SVM, Figure 7) perform better. However, the difference in DS is determined by only a few seizures, so these results are quite noisy, and one should not draw strong conclusions based on these results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 December 2020 doi:10.20944/preprints202012.0527.v1

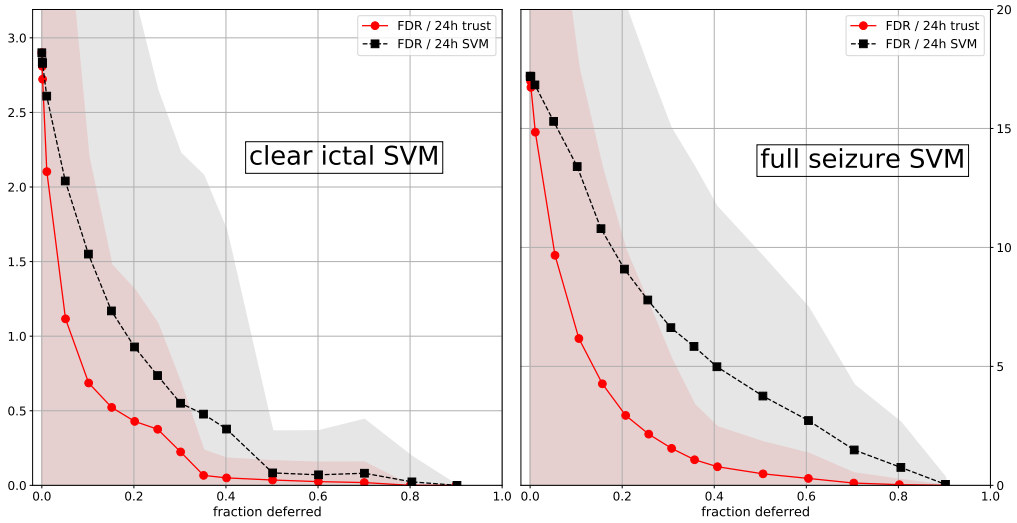

Figure 10. Average FDR / 24 hours as a function of the fraction of the data that is deferred to a human annotator, for the CI and FS SVM ($p_{\text{low}} = 5$). The standard deviation is shown as a shaded area. In contrast to our main approach, the segments that contain seizure flags are not automatically the first to be deferred. Segments are deferred using the SVM confidences (SVM) and trust scores (trust) from a trust model trained on the FS labels (for both cases).

4. Discussion

To appreciate how good deferring segments to a human annotator works, note that for the baseline strategy of deferring random segments the average performance stay constant and become more noisy. Compared to this scenario the scaling of the performances is good: the average performances increase monotonically and the standard deviations decrease. We investigated some basic deferral strategies: Taking the average of the least-confident percentage p_{low} of the 2-second segments in the 5-minute segments; using SVM confidences or trust scores; and taking the same or a different percentage of the data to defer per patient (discussed in the supplementary material). These strategies already showed significant variation in performance. Perfect performance can be reached when deferring 50% of the data. To achieve clinical applicability, this should be reduced to around 5% to 10%. Our results indicate that more sophisticated deferral schemes could further improve our results. In our application, the SVM is not aware that it can defer data. Which data to defer is decided after training the SVM. One can train a classifier and a rejector simultaneously [26,30–32], which can improve both the classifier and the deferral decisions. Our current strategy consists of selecting 5-minute segments, and taking the average of the confidences of the 2-second segments included in this 5-minute segment. Making the length of the segments variable, and making the deferral choice based on more detailed properties then the average confidence, could be beneficial. Better classification models and better confidence estimates are also a relevant research direction. Finally, more or multi-modal data would likely improve the performance.

For the investigated dataset, LTF is a valuable strategy to improve the performance of an already trained classifier. It is relatively straightforward to implement, since only the "percentile of segments to filter" hyper-parameter needs to be optimized. If one fits a model and afterwards performs LTF, the likelihood of introducing over-fitting is therefore small. In the nested cross-validation scheme of this article, the probability of over-fitting is negligible. Calculating the trust score for a 2-second segment takes 0.0016 seconds on a laptop (dual core i5-6200 CPU with 16 GB RAM), so LTF could be applied in online seizure detection. A mechanism similar to LTF could be achieved by attention networks [41]. These are neural networks that learn to weight the input, depending on its relevance. Attention networks have already been applied to seizure detection [16].

 doi:10.20944/preprints202012.0527.v1

Note that since we remove the lowest-trusted percentiles for both classification with a deferral option and LTF, we are interested in a good uncertainty ranking of the predictions, which is not necessarily the same as a good calibration. A well-calibrated model returns probabilities that reflect the likelihood of its predictions. Deep neural networks seem to return good uncertainty rankings [28,37], even though they are not necessarily well calibrated [37]. There are empirical studies on what methods are optimal for calibrating machine learning classifiers [42,43]. SVMs that are calibrated by Platt scaling perform moderately well, whereas random forests and small neural networks perform better. It is, however, unclear if the results from such studies can be directly translated to seizure detection, for which there is a very large class imbalance. There is, furthermore, evidence that the optimal calibration method can change depending on the dataset. It was found that datasets with labels that contain some inherent uncertainty (i.e., where experts would disagree on some labels) benefit from different methods compared to datasets without that extra uncertainty [44]. Given these issues, progress on achieving models with good uncertainty rankings will likely be made by empirical studies specifically focused on the seizure detection task. As already mentioned, another option is to investigate algorithms with an explicit reject (defer) option [30–32].

Trust scores outperform SVM confidences, with the exception of the DS for classification with a deferral option, where they perform similarly. The main claim of the article that introduces trust scores is that they produce better uncertainty rankings compared to the classifier itself, at least on low to mid-dimensional datasets [28]. The results on this seizure detection task corroborate this claim. We made an attempt to understand of what causes low trust scores in the EEG data (e.g. they are mostly caused by measurement noise), but this proved to be difficult.

It is an interesting question how model confidences and trust scores perform on non-EEG data such as electrocardiogram, photoplethysmography, electromyogram, or accelerometry data [8,9,45], both for classification with a deferral option and LTF. In a multi-modal setting, one can investigate if confidence measures can be used to detect which modality works best for detecting seizures for a given patient. The use of different modalities is of interest in some types of seizures since they can be better detected with the use of an alternative biosignal (e.g., EMG in tonic clonic seizures).

5. Conclusions

We have investigated two applications of seizure detection where the classifier has the option not to make a decision. The dataset under study consists of EEG measurements from four behind-the-ear sensors on 54 epileptic patients with focal onset seizures [21]. We expect similar results for other types of seizures (e.g., absence seizures), since the proposed method is being used as a post-processing tool and the type of seizure will not affect (at least not significantly) the conclusions reached in this study. Behind-the-ear measurements can be used for long-term home monitoring outside the hospital. Support vector machines classifiers were already developed in previous work [21]. Prediction confidences are determined by temperature scaling [37] of the SVM output and trust scores [28], which can be calculated independently from the classifier.

In the first application, we investigate the performance gain in the case that part of the data is deferred to a human annotator, who is assumed to annotate perfectly. For both models, a detection sensitivity of approximately 90% (99%) can be achieved when deferring around 10% (40%) of the data. Perfect performance can be reached after deferring 50% of the data. Our results indicate that better deferral strategies, improved classifiers, and better confidence measures could provide further improvements.

In the second application we show that a common modelling strategy for EEG data, where predictions from several short EEG segments are used to obtain a final prediction [9,20,21,33], can be improved by filtering out untrustworthy segments. The false detection rate shows a relative decrease between 21% and 43%, and the detection sensitivity shows a small increase or decrease. Both the positive predictive value and

 doi:10.20944/preprints202012.0527.v1

F1-score improve considerably. Filtering only works with trust scores. It does not work with the confidences calculated from the SVM output. This corroborates the results from [28]. Since only one hyper-parameter needs to be optimized, these results suggest that this approach is a relatively straightforward way to improve the performance of a pre-trained classifier, without introducing over-fitting.

Supplementary Materials: The following are available online at https://www.mdpi.com/1424-8220/1/1/0/s1, Figure S1: Average performance on all patients (no cross-validation) as a function of the percent of 2-second segments that are filtered (% removed), for the first rule investigated, Figure S2: Average performance on all patients (no cross-validation) as a function of the percent of 2-second segments that are filtered (% removed), for the second rule investigated, Figure S3: Average performance as a function of the fraction of the data that is deferred to a human annotator, for the CI SVM ($p_{\text{low}} = 5$), with trust scores (trust) from a trust model trained on the CI labels, Figure S4: Average performance as a function of the fraction of the data that is deferred to a human annotator, for the FS SVM ($p_{\text{low}} = 5$), with trust scores (trust) from a trust model trained on the CI labels, Figure S5: Average number and average length (minutes) of the deferred segments, as a function of the fraction of the data that is deferred to a human annotator, for the FS SVM ($p_{\text{low}} = 5$), Figure S6: Average FDR as a function of the fraction of the data that is deferred to a human annotator, for the CI SVM, for different p_{low}. The segments that contain seizure flags are not deferred at the start, Figure S7: Average performance as a function of the fraction of the data that is deferred to a human annotator, for the CI SVM ($p_{\text{low}} = 5$). The number of segments that is deferred is patient independent or patient dependent, Figure S8: Average performance as a function of the fraction of the data that is deferred to a human annotator, for the CI SVM ($p_{\text{low}} = 5$), with and without LTF, Figure S9: Average performance as a function of the fraction of the data that is deferred to a human annotator, for the FS SVM ($p_{\text{low}} = 5$), with and without LTF, Figure S10: A visualisation of a new detection after low-trust filtering for the CI SVM, Table S1: Mean (standard deviation) performance of: CI SVM, LTF with trust models trained on the FS (trust FS) labels, with the percentage of segments filtered either patient independent or patient dependent, Table S2: Mean (standard deviation) performance of: CI SVM, LTF with trust models trained on the CI (trust CI) labels, with the percentage of segments filtered either patient independent or patient dependent, Table S3: Mean (standard deviation) performance of: CI SVM, LTF with the confidences of this SVM (CI SVM conf.), with the percentage of segments filtered either patient independent or patient dependent.

Author Contributions: Conceptualization, Thijs Becker; Data curation, Kaat Vandecasteele; Formal analysis, Thijs Becker and Kaat Vandecasteele; Funding acquisition, Dirk Valkenborg and Sabine Van Huffel; Investigation, Thijs Becker and Kaat Vandecasteele; Methodology, Thijs Becker, Kaat Vandecasteele, Christos Chatzichristos, Dirk Valkenborg, Sabine Van Huffel and Maarten De Vos; Project administration, Christos Chatzichristos, Dirk Valkenborg, Sabine Van Huffel and Maarten De Vos; Resources, Dirk Valkenborg, Sabine Van Huffel and Maarten De Vos; Software, Thijs Becker and Kaat Vandecasteele; Supervision, Christos Chatzichristos, Dirk Valkenborg, Sabine Van Huffel and Maarten De Vos; Validation, Thijs Becker and Kaat Vandecasteele; Visualization, Thijs Becker; Writing – original draft, Thijs Becker; Writing – review & editing, Thijs Becker, Kaat Vandecasteele, Christos Chatzichristos, Wim Van Paesschen, Dirk Valkenborg, Sabine Van Huffel and Maarten De Vos. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. K.V, C.C., S.V.H., M.D.V. are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium. Funding received from EIT 19263 – SeizeIT2: Discreet Personalized Epileptic Seizure Detection Device.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the ethical commission research of UZ Leuven (S-63900 on 11 May, 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from W.V.P. or M.D.V. The data are not publicly available due to privacy reasons.

Acknowledgments: SeizeIT is a project realized in collaboration with IMEC. Project partners are KU Leuven, UCB Pharma, Byteflies and Pilipili, with project support from VLAIO (Flanders Innovation and Entrepreneurship) and Innoviris. ICON: HBC.2016.0167 SeizeIT.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CI	clear ictal
EEG	electroencephalographic
FDR	false detection rate
FN	false negative
FP	false positive
FS	full seizure
KRCC	Kendall rank correlation coefficient
LTF	low-trust filtering
PPV	positive predictive value
SVM	support vector machine
TP	true positive

References

1. Fiest, K.M.; Sauro, K.M.; Wiebe, S.; Patten, S.B.; Kwon, C.S.; Dykeman, J.; Pringsheim, T.; Lorenzetti, D.L.; Jetté, N. Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. *Neurology* **2017**, *88*, 296–303.
2. French, J.A. Refractory Epilepsy: Clinical Overview. *Epilepsia* **2007**, *48*, 3–7, [https://onlinelibrary.wiley 1167.2007.00992.x]. doi:10.1111/j.1528-1167.2007.00992.x.
3. Elger, C.E.; Mormann, F. Seizure prediction and documentation—two important problems. *The Lancet Neurology* **2013**, *6*, 531–532.
4. Brunnhuber, F.; Slater, J.; Goyal, S.; Amin, D.; Thorvardsson, G.; Freestone, D.R.; Richardson, M.P. Past, Present and Future of Home video-electroencephalographic telemetry: A review of the development of in-home video-electroencephalographic recordings. *Epilepsia*, *n/a*, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/epi.16578]. doi:10.1111/epi.16578.
5. Fisher, R.S.; Blum, D.E.; DiVentura, B.; Vannest, J.; Hixson, J.D.; Moss, R.; Herman, S.T.; Fureman, B.E.; French, J.A. Seizure diaries for clinical research and practice: Limitations and future prospects. *Epilepsy & Behavior* **2012**, *24*, 304 – 310. doi:https://doi.org/10.1016/j.yebeh.2012.04.128
6. Baumgartner, C.; Koren, J.P. Seizure detection using scalp-EEG. *Epilepsia* **2018**, *59*, 14–22, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/epi.14052]. doi:10.1111/epi.14052.
7. Beniczky, S.; Karoly, P.; Nurse, E.; Ryvlin, P.; Cook, M. Machine learning and wearable devices of the future. *Epilepsia*, *n/a*, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/epi.16555]. doi:10.1111/epi.16555.
8. Leijten, F.S.S.; the Dutch TeleEpilepsy Consortium. Multimodal seizure detection: A review. *Epilepsia* **2018**, *59*, 42–47, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/epi.14047]. doi: https://doi.org/10.1111/epi.14047.
9. De Cooman, T.; Varon, C.; Van de Vel, A.; Ceulemans, B.; Lagae, L.; Van Huffel, S. Comparison and combination of electrocardiogram, electromyogram and accelerometry for tonic-clonic seizure detection in children. 2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), 2018, pp. 438–441. doi:10.1109/BHI.2018.8333462.
10. EIT Health. SeizeIT2. https://eithealth.eu/project/seizeit2/, 2018. Accessed: 2020-12-14.
11. Kim, T.; Nguyen, P.; Pham, N.; Bui, N.; Truong, H.; Ha, S.; Vu, T. Epileptic Seizure Detection and Experimental Treatment: A Review. *Frontiers in Neurology* **2020**, *11*, 701. doi: 10.3389/fneur.2020.00701.
12. Siddiqui, M.K.; Morales-Menendez, R.; Huang, X.; Hussain, N. A review of epileptic seizure detection using machine learning classifiers. *Brain Informatics* **2020**, *7*, 1–18.
13. Shoeibi, A.; Ghassemi, N.; Khodatars, M.; Jafari, M.; Hussain, S.; Alizadehsani, R.; Moridian, P.; Khosravi, A.; Hosseini-Nejad, H.; Rouhani, M.; Zare, A.; Khadem, A.; Nahavandi, S.;

Atiya, A.F.; Acharya, U.R. Epileptic seizure detection using deep learning techniques: A Review, 2020, [arXiv:cs.LG/2007.01276].
14. González Otárula, K.A.; Mikhaeil-Demo, Y.; Bachman, E.M.; Balaguera, P.; Schuele, S. Automated seizure detection accuracy for ambulatory EEG recordings. *Neurology* **2019**, *92*, e1540–e1546, [https://n.neurology.org/content/92/14/e1540.full.pdf]. doi:10.1212/WNL.0000000000007237.
15. Shah, V.; von Weltin, E.; Lopez, S.; McHugh, J.R.; Veloso, L.; Golmohammadi, M.; Obeid, I.; Picone, J. The Temple University Hospital Seizure Detection Corpus. *Frontiers in Neuroinformatics* **2018**, *12*, 83. doi:10.3389/fninf.2018.00083.
16. Chatzichristos, C.; Dan, J.; Narayanan, A.; Seeuws, N.; Vandecasteele, K.; De Vos, M.; Bertrand, A.; Van Huffel, S. Epileptic Seizure Detection in EEG via Fusion of Multi-View Attention-Gated U-net Deep Neural Networks. Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE, 2020.
17. SeizeIT1. https://www.imec-int.com/en/what-we-offer/research-portfolio/seizeit, 2016. Accessed: 2020-12-14.
18. Boeckx, S.; van Paesschen, W.; Bonte, B.; Dan, J. Live Demonstration: SeizeIT - A wearable multimodal epileptic seizure detection device. 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2018, pp. 1–1. doi:10.1109/BIOCAS.2018.8584738.
19. Zibrandtsen, I.; Kidmose, P.; Christensen, C.; Kjaer, T. Ear-EEG detects ictal and interictal abnormalities in focal and generalized epilepsy – A comparison with scalp EEG monitoring. *Clinical Neurophysiology* **2017**, *128*, 2454 – 2461. doi:https://doi.org/10.1016/j.clinph.2017.09.115.
20. Gu, Y.; Cleeren, E.; Dan, J.; Claes, K.; Van Paesschen, W.; Van Huffel, S.; Hunyadi, B. Comparison between Scalp EEG and Behind-the-Ear EEG for Development of a Wearable Seizure Detection System for Patients with Focal Epilepsy. *Sensors* **2018**, *18*. doi:10.3390/s18010029.
21. Vandecasteele, K.; De Cooman, T.; Dan, J.; Cleeren, E.; Van Huffel, S.; Hunyadi, B.; Van Paesschen, W. Visual seizure annotation and automated seizure detection using behind-the-ear electroencephalographic channels. *Epilepsia* **2020**, *61*, 766–775, [https://onlinelibrary.wiley.com/doi/p doi:10.1111/epi.16470.
22. You, S.; Cho, B.H.; Yook, S.; Kim, J.Y.; Shon, Y.M.; Seo, D.W.; Kim, I.Y. Unsupervised automatic seizure detection for focal-onset seizures recorded with behind-the-ear EEG using an anomaly-detecting generative adversarial network. *Computer Methods and Programs in Biomedicine* **2020**, *193*, 105472. doi:https://doi.org/10.1016/j.cmpb.2020.105472.
23. Clarke, S.; Karoly, P.J.; Nurse, E.; Seneviratne, U.; Taylor, J.; Knight-Sadler, R.; Kerr, R.; Moore, B.; Hennessy, P.; Mendis, D.; Lim, C.; Miles, J.; Cook, M.; Freestone, D.R.; D'Souza, W. Computer-assisted EEG diagnostic review for idiopathic generalized epilepsy. *Epilepsy & Behavior* **2019**, p. 106556. doi:https://doi.org/10.1016/j.yebeh.2019.106556.
24. Chow, C.K. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers* **1957**, *EC-6*, 247–254. doi:10.1109/TEC.1957.5222035.
25. Chow, C. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* **1970**, *16*, 41–46. doi:10.1109/TIT.1970.1054406.
26. Bartlett, P.L.; Wegkamp, M.H. Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research* **2008**, *9*, 1823–1840.
27. Cortes, C.; DeSalvo, G.; Mohri, M. Learning with Rejection. Algorithmic Learning Theory; Ortner, R.; Simon, H.U.; Zilles, S., Eds.; Springer International Publishing: Cham, 2016; pp. 67–82.
28. Jiang, H.; Kim, B.; Guan, M.; Gupta, M. To Trust Or Not To Trust A Classifier. In *Advances in Neural Information Processing Systems 31*; Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; Garnett, R., Eds.; Curran Associates, Inc., 2018; pp. 5541–5552.
29. Madras, D.; Pitassi, T.; Zemel, R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems 31*; Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; Garnett, R., Eds.; Curran Associates, Inc., 2018; pp. 6147–6157.
30. Raghu, M.; Blumer, K.; Corrado, G.; Kleinberg, J.; Obermeyer, Z.; Mullainathan, S. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort, 2019, [arXiv:cs.CV/1903.12220].
31. De, A.; Okati, N.; Zarezade, A.; Gomez-Rodriguez, M. Classification Under Human Assistance, 2020, [arXiv:stat.ML/2006.11845].
32. Mozannar, H.; Sontag, D. Consistent Estimators for Learning to Defer to an Expert, 2020, [arXiv:cs.LG/2006.01862].
33. Kuhlmann, L.; Karoly, P.; Freestone, D.R.; Brinkmann, B.H.; Temko, A.; Barachant, A.; Li, F.; Titericz, Gilberto, J.; Lang, B.W.; Lavery, D.; Roman, K.; Broadhead, D.; Dobson, S.; Jones, G.;

Tang, Q.; Ivanenko, I.; Panichev, O.; Proix, T.; Náhlík, M.; Grunberg, D.B.; Reuben, C.; Worrell, G.; Litt, B.; Liley, D.T.J.; Grayden, D.B.; Cook, M.J. Epilepsyecosystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG. *Brain* **2018**, *141*, 2619–2630, [https://academic.oup.com/brain/article-pdf/141/9/2619/25590596/awy210.pdf]. doi:10.1093/brain/awy210.

34. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
35. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S.J.; Brett, M.; Wilson, J.; Jarrod Millman, K.; Mayorov, N.; Nelson, A.R.J.; Jones, E.; Kern, R.; Larson, E.; Carey, C.; Polat, l.; Feng, Y.; Moore, E.W.; Vand erPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E.A.; Harris, C.R.; Archibald, A.M.; Ribeiro, A.H.; Pedregosa, F.; van Mulbregt, P.; Contributors, S... SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**.
36. Debener, S.; Emkes, R.; De Vos, M.; Bleichner, M. Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear. *Scientific reports* **2015**, *5*, 16743.
37. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. Proceedings of Machine Learning Research; Precup, D.; Teh, Y.W., Eds.; PMLR: International Convention Centre, Sydney, Australia, 2017; Vol. 70, *Proceedings of Machine Learning Research*, pp. 1321–1330.
38. Platt, J.C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in large margin classifiers. MIT Press, 1999, pp. 61–74.
39. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **1974**, *36*, 111–147.
40. Wainer, J.; Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications, 2018, [arXiv:cs.LG/1809.09446].
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30, pp. 5998–6008.
42. Niculescu-Mizil, A.; Caruana, R. Predicting Good Probabilities with Supervised Learning. Proceedings of the 22nd International Conference on Machine Learning; Association for Computing Machinery: New York, NY, USA, 2005; ICML '05, p. 625–632. doi:10.1145/1102351.1102430.
43. Caruana, R.; Karampatziakis, N.; Yessenalina, A. An Empirical Evaluation of Supervised Learning in High Dimensions. Proceedings of the 25th International Conference on Machine Learning; Association for Computing Machinery: New York, NY, USA, 2008; ICML '08, p. 96–103. doi:10.1145/1390156.1390169.
44. Rousseau, A.J.; Becker, T.; Bertels, J.; Blaschko, M.B.; Valkenborg, D. Post Training Uncertainty Calibration of Deep Networks For Medical Image Segmentation, 2020, [arXiv:eess.IV/2010.14290].
45. Vandecasteele, K.; De Cooman, T.; Gu, Y.; Cleeren, E.; Claes, K.; Van Paesschen, W.; Van Huffel, S.; Hunyadi, B. Automated Epileptic Seizure Detection Based on Wearable ECG and PPG in a Hospital Environment. *Sensors* **2017**, *17*. doi:10.3390/s17102338.