Stevenson, Nathan; Tapani, Karoliina; Vanhatalo, Sampsa

Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert

Published in:
2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2019

DOI:
10.1109/EMBC.2019.8857367

Published: 01/07/2019

Document Version
Peer reviewed version

Please cite the original version:
Stevenson, N., Tapani, K., & Vanhatalo, S. (2019). Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2019* (pp. 5991-5994). [8857367] IEEE. https://doi.org/10.1109/EMBC.2019.8857367

Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert*

Nathan Stevenson[1], Karoliina Tapani[2], and Sampsa Vanhatalo[3]

***Abstract*— Neonatal EEG seizure detection algorithms (NSDAs) have an upper bound of performance related to the agreement between visual interpretation of human experts. No published algorithms have reported performance that has reached this upper bound. In this paper, we combined two recently developed NSDAs in order to improve detection performance. An outlier detection stage was also added to improve robustness in the presence of unseen data. A large database of EEG from 79 term infants labeled by three independent human experts was used to develop and test the sufficiency of the hybrid NSDA. The inter-observer agreement (IOA) between experts was high (κ = 0.757, 95%CI: 0.665-0.836, n=79). The area under the receiver operator characteristic of the NSDA compared to the consensus annotation of the human experts was 0.952 (95%CI: 0.0927-0.971). The IOA of seizure detection between the NSDA and human experts was not significantly less than the IOA among human experts ($\Delta\kappa$ = 0.022, 95%CI: -0.020 to 0.072) and was further improved by increasing the minimum seizure duration from 10s to 30s ($\Delta\kappa$ = -0.002, 95%CI: -0.073 to 0.055). Automated methods of neonatal EEG seizure detection have sufficient accuracy to replace human interpretation, potentially, providing reliable interpretations for clinicians in the neonatal intensive care unit.**

I. Introduction

Neonatal seizures are observed in 1-5 per 1000 live births and are urgently treated with a range of anti-epileptic drugs (AED) in order to minimize the overall seizure burden (the accumulated seizure duration) in a neonate [1]. Seizures are treated as clinical studies have shown that neonatal seizures are associated with abnormal neurodevelopmental outcome [3]. The effectiveness of any potential treatment is closely linked with seizure detection [2].

The detection of neonatal seizures is not trivial as clinical signs are often absent or suppressed by medications [4]. The current gold standard for neonatal seizure detection is visual interpretation of electrical activity of the brain, recorded by the EEG complemented by video, ECG and respiration, by the human expert [5]. This expertise is, however, not available on demand to clinicians for optimal AED targeting and monitoring of AED effectiveness. A relatively cost-effective solution to this problem would be to embed automated neonatal EEG seizure detection algorithms (NSDA) into EEG machines to provide the attending clinician with a signal trend that indicates the presence of neonatal seizures.

There are several NSDAs reported in the recent literature [6], [7]. These methods have shown increasing performance over the years and their incorporation into standard monitoring practice has been shown to improve seizure recognition in the NICU [2]. The performance of these NSDA algorithms, however, has not yet reached the benchmark of visual interpretation of the EEG by the human expert (a benchmark that can be quantified using measures of inter-observer agreement, IOA). There are several potential sources of sub-optimal performance from features, algorithm structure, and classification algorithm. There is also the question of generalisation of the developmental dataset to a larger population of EEG recordings across varied sites, underlying etiologies, demographics and EEG technologies. More importantly, researchers use performance metrics based on the assumption of an immutable gold standard of annotation. This is not the case, and NSDAs are not required to achieve 100% performance to be sufficiently accurate, rather they should be benchmarked against the IOA between human experts.

In this paper we develop a hybrid NSDA based on the combination of two recently developed NSDAs in an attempt to improve seizure detection performance. We also add an outlier detection post-processing stage to improve the robustness of the NSDA to unseen data. We analyse NSDA performance in terms of IOA and show that replacing the annotation of one human expert with the hybrid NSDA output does not significantly reduce IOA of an 'all human' annotation resulting in a sufficiently accurate NSDA.

II. Data

We used a publicly available dataset of neonatal EEG recordings and annotations [8]. In brief, EEGs were collected from 79 neonates admitted to the neonatal intensive care unit (NICU) at the Helsinki University Central Hospital between 2010 and 2014. The EEG was recorded with a NicOne EEG amplifier (sampling frequency of 256 Hz; Natus, USA) and EEG caps (sintered Ag/AgCl electrodes; Waveguard, ANT-Neuro, Germany). These caps contained 19 electrodes positioned using the international 10–20 standard [8]. A bipolar montage was constructed from these electrodes and used for seizure detection (Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz, Cz-Pz). Each EEG recording was annotated for seizures independently by three experts who annotated the start and end of a seizure

*This work was supported by Academy of Finland (276523/288220) and a Marie Skłodowska Curie Fellowship (H2020-MCSA-IF-656131).

[1]Nathan Stevenson is with QIMR Berghofer Medical Research Institute, 300 Herston Rd, 4006, Brisbane, Australia nathan.stevenson@qimrberghofer.edu.au

[2] Karoliina Tapani is with Aalto University, Otakaari 1, 02150, Espoo, Finland karoliina.tapani@aalto.fi

[3] Sampsa Vanhatalo is with the Department of Neurophysiology, Helsinki University Hospital and the University of Helsinki, Yliopistonkatu 4, 00100 Helsinki, Finland sampsa.vanhatalo@helsinki.fi

taking into account all EEG channels. Data collection was approved by the Ethics Committee of the Helsinki University Children's Hospital. All patient identifying information was removed from each recording before annotations.

III. Hybrid Seizure Detection Algorithm

The hybrid algorithm combined the feature set developed in [7] with the output of the convolutional neural network (CNN) developed in [6], using a kernel support vector machine (SVM) to form a decision output (see Fig. 1). An outlier detection stage was also added to the post-processing stage to ensure that only EEG epochs that fit within the span of the training feature space could be detected as seizure.

For the feature extraction stage, the EEG for each channel was filtered with a high pass filter (4th order Butterworth, cutoff of 0.5Hz), resampled to 64Hz (with an anti-aliasing filter) and then segmented into 16 second epochs (discrete signal length of 1024 samples) with an overlap of 12 seconds. A feature set of 22 features was extracted from each EEG epoch. These features attempt to represent the time-varying periodicity in the EEG signal and are supplemented with several summary measures of the spectrum and nonlinear energy output [8].

For the CNN stage, the EEG for each channel was filtered with a bandpass pass filter (8th order Chebyshev Type 2, cutoff of [0.5,16]Hz), resampled to 32Hz (with an anti-aliasing filter) and then segmented into 16 second epochs (discrete signal length of 512 samples) with an overlap of 12 seconds.

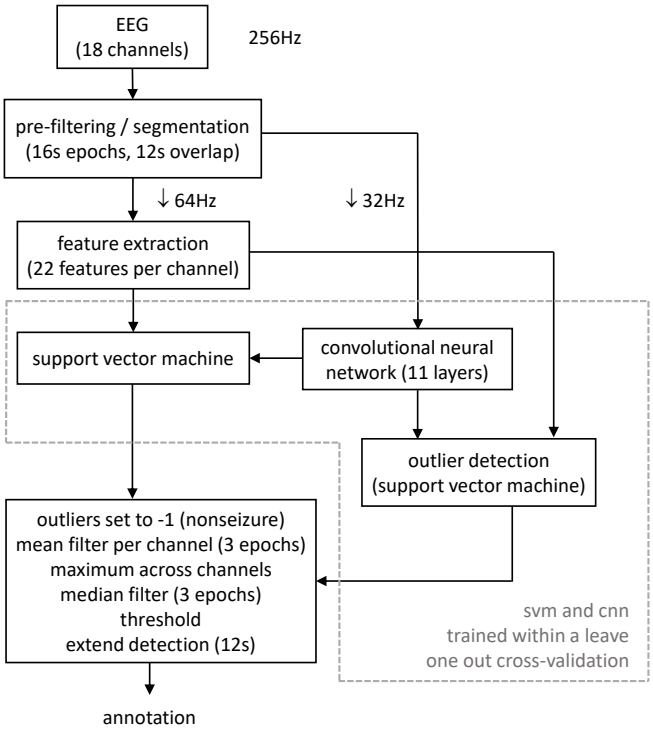

Fig. 1. The structure of the NSDA

IV. Training and Testing

Training and testing was performed within a leave-one-subject out cross-validation (LOSO). Training data consisted of a 96103 epochs of EEG (19339 seizure, 76764 non-seizure). The average training set size per iteration of the LOSO was 94886 epochs of EEG (19094 seizure, 75792 nonseizure). SVMs were trained using Matlab (2018b). A radial basis kernel was used and SVM parameters were optimized within each training iteration using 5-fold cross-validation and Bayesian optimisation. For outlier detection, the bias was set such that 5% of the data was deemed to be outliers. The CNN was also implemented in Matlab (2018b). The structure of the deep network was almost identical to that used in [6]: 11 convolutional layers, rectified linear unit, batch normalization with average pooling. The only modifications were a larger input size (due to the use of 16s epochs rather than 8s epochs) and replacement of the final global average pooling layer with a fully connected layer. The CNN was trained using stochastic gradient descent with momentum (0.9), batch size of 1024 and epoch length of 70 on a NVIDIA GTX 1080 GPU. The intial learn rate was 0.1 reduced by a factor of 0.2 every 20 epochs.

The SVM output (see eqn (80) in [12]) was converted into a single binary decision via several post processing stages. The SVM output per channel was convolved with a moving average filter of 3 samples in length (12s). The maximum value across 18 channels was taken and then a median filter (3 samples, 12s) was applied to this output. A threshold was used to form a final binary decision (a single threshold was used for all iterations of the LOSO). The binary output was then extended in order to account for the overlapping procedure of EEG segmentation. The threshold of detection and extension period were set to maximize the IOA between the human experts and NSDA output. The output of the SVM outlier detection stage was incorporated into the first stage of post-processing. The SVM output per channel was set to -1 (nonseizure) when the EEG epoch was considered as an outlier. The outlier was detected using a single threshold of the outlier SVM output (threshold selected to maximize IOA between human experts and NSDA).

Algorithm performance was assessed using measures of IOA (Fleiss' kappa, [9]). When calculating the IOA, all annotations were concatenated (linked) together to form a single annotation for the entire dataset. IOA was calculated on a second by second basis. To assess the sufficiency of the proposed NSDA, the IOA was calculated by substituting the annotation of a single human observer with the output of the NSDA. The difference between the IOA of the 'all human' annotation and the 'human/NSDA' annotation was then evaluated with a bootstrap (1000 resamplings). If the 95% confidence interval (95%CI) contained 0 for at least one substituted human expert, then the NSDA annotation was assumed to be sufficiently accurate to replace a human expert.

The area under the receiver operator characteristic curve (AUC), seizure detection rate (SDR) and false detections per

TABLE I

PERFORMANCE OF THE PROPOSED NEONATAL SEIZURE DETECTION ALGORITHMS (NSDA) EVALUATED ON CONSENSUS RECORDINGS (PERIODS OF NON-CONSENSUS IGNORED). PERFORMANCE MEASURES WERE CALCULATED ON EACH NEONATE AND THEN SUMMARIZED ACROSS THE COHORT (PRESENTED AS MEDIAN [INTERQUARTILE RANGE] AND DENOTED WITH A SUBSCRIPT S) AND ON A CONCATENATED RECORDINGS (PRESENTED AS VALUE [95% CONFIDENCE INTERVAL] AND DENOTED WITH A SUBSCRIPT C). SEE THE TRAINING AND TESTING SECTION FOR DEFINITIONS OF NSDA ACRONYMS. AUC IS AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC, SDR IS SEIZURE DETECTION RATE, AND FD/H IS FALSE DETECTIONS PER HOUR. NOTE, THAT AUC AND SDR ARE SUMMARIZED OVER 39 NEONATES (ONLY NEONATES WITH CONSENSUS SEIZURE) AND FD/H ARE SUMMARIZED OVER ALL 79 NEONATES. WHEN RESTRICTING SEIZURES TO LESS THAN 30S, AUC AND SDR ARE SUMMARIZED OVER ONLY 36 NEONATES AS 3 NEONATES DID NOT HAVE ANY SEIZURE OVER 30S.

	AUC_s	SDR_s	FD/h_s	AUC_c	SDR_c	FD/h_c
SVM	0.991 [0.949-0.999]	100% [73-100%]	0.8 [0-2.4]	0.966 [0.942-0.983]	79% [66-90%]	2.1 [1.3-3.0]
CNN	0.991 [0.964-0.999]	100% [81-100%]	1.2 [0-4.0]	0.963 [0.937-0.984]	82% [71-91%]	2.6 [1.8-3.6]
CNN+SVM	0.994 [0.953-0.999]	100% [69-100%]	0 [0-1.7]	0.967 [0.949-0.984]	76% [61-89%]	1.5 [0.8-2.3]
CNN+SVM+OUT	0.977 [0.884-0.997]	100% [54-100%]	0 [0-1.8]	0.952 [0.927-0.971]	71% [55-86%]	1.0 [0.6-1.7]
CNN+SVM+OUT (30s)	0.979 [0.920-0.996]	100% [78-100%]	0 [0-0]	0.953 [0.930-0.972]	85% [75-93%]	0.5 [0.2-0.9]

hour (FD/h) were also calculated per neonate. The sensitivity and specificity used to calculate the AUC were based on a temporal assessment; a true positive was correct detection of 1s of seizure and a true negative was correct detection of 1s of non-seizure. For the event based assessment (seizure detection rate, false detection eper hour), any temporal overlap between an NSDA defined seizure and the annotation of the human experts was considered as a true seizure detection.

Several systems were evaluated in order to assess improvements due to additional stages in the NSDA: SVM - support vector machine with feature set (22 features), CNN - convolutional neural network (11 convolutional layers), SVM+CNN - support vector machine with feature set and CNN output as a feature (23 features), SVM+CNN+OUT - support vector machine with feature set and CNN output (23 features), and post-processing based on outlier detection. As the IOA is known to be dependent on seizure duration with higher agreement for longer duration seizures [9], we also evaluated the performance of a NSDA with the minimum seizure duration increased from 10s to 30s (events less than 30s are redefined as nonseizure).

V. RESULTS AND DISCUSSION

There was little difference between SVM, CNN and combined SVM/CNN based NSDAs when evaluated with traditional methods of performance assessment (see Table I). For a decision threshold that maximized the IOA, the SVM based method had slightly less false detections but a lower seizure detection rate. In this case, performance measures on the concatentated recordings are more useful to assess performance as they do not reach the upper bound of the measurement. The inclusion of an outlier detection stage results in a decrease in AUC and SDR, but less FD/h. Increasing the minimum seizure duration from 10s to 30s, provides the largest increase in performance with the highest SDR and lowest FD/h.

The use of consensus annotations is biased towards higher performance; the TDR and FD/h measures when calculated on the concatenated recordings of individual experts, and then averaged, for the SVM+CNN+OUT NSDA was 53% (95%CI: 41-65%) and 1.5/h (95%CI: 0.8-2.3/h), respectively, as opposed to the consensus results of 71% [55-86%] and 1.0/h [0.6-1.7/h], respectively.

In terms of IOA, all NSDAs achieved the benchmark performance of not significantly reducing the IOA when substituting a single human expert (see Table II and Fig. 2. All NSDAs had a 95%CI of differences between the inter-observer agreement between an all human and human/NSDA annotation that contains 0 for at least one human/SDA annotation. Incremental improvements were seen when combining the feature set and CNN output and adding an outlier detection post-processing stage. The change in NSDA with outlier detection from inferior when assessed using AUC to superior when assessed using IOA was most likely due to the selection of the outlier threshold to maximize IOA. This shows that IOA and AUC are not perfectly correlated as measures of NSDA performance and that higher IOA appear associated with lower FD/h. We also show that by simply reducing the EEG epoch length (from 32s to 16s) improved the performance of our original NSDA (22 features SVM) [7]. IOA between all human experts was 0.757 (95%CI: 0.665-0.836) and with a minimum seizure duration of 30s IOA was 0.754 (95%CI: 0.654-0.838).

The largest improvement in IOA was seen when increasing the minimum seizure duration from 10s to 30s. The increase of minimum seizure duration does not considerably alter the measured seizure burden: neonates with consensus seizure decreases from 39 to 36 and the seizure burden within these consensus neonates (averaged across experts) decreases from a median of 15.4 mins (IQR: 6.5-30.9 mins, n=39) to a median of 15.3 mins (IQR: 6.5-27.9 mins, n=36).

We claim that a 95%CI of differences between the IOA between an all human and human/NSDA annotation that contains 0 for at least one human/SDA annotation is sufficient to define an NSDA annotation that is indistinguishable from the human expert. This is optimistic and may not be strict enough to satisfy clinicians or regulatory authorities. This does not invalidate the assessment methods; rather stricter conditions, such as an 80%CI for all human/NSDA combinations, may be required. While the sufficiency of proposed NSDA performance may not be definitive and only

TABLE II

THE AGREEMENT (κ) BETWEEN COMPOSITE NSDA/HUMAN ANNOTATIONS AND THE 95% CONFIDENCE INTERVAL OF DIFFERENCES IN κ BETWEEN THE ALL HUMAN ANNOTATION (3 EXPERTS; E1, E2 AND E3) AND A COMPOSITE NSDA/HUMAN ANNOTATION (NSDA + 2 HUMAN EXPERT). THE RESULTS ARE PRESENTED AS κ OF HYBRID NSDA/HUMAN EXPERT (95%CI OF $\Delta\kappa$ BETWEEN HYBRID NSDA/HUMAN EXPERTS ANNOTATION AND THE ALL HUMAN ANNOTATION)

	replace E1 with NSDA	replace E2 with NSDA	replace E3 with NSDA
SVM	0.655 (0.041-0.176)	0.700 (-0.012-0.146)	0.640 (0.065-0.186)
CNN	0.670 (0.036-0.138)	0.716 (-0.018-0.106)	0.660 (0.053-0.143)
SVM+CNN	0.676 (0.021-0.150)	0.722 (-0.028-0.116)	0.659 (0.051-0.161)
SVM+CNN+OUT	0.690 (0.030-0.109)	0.735 (-0.020-0.072)	0.670 (0.056-0.122)
SVM+CNN+OUT (30s)	0.695 (0.021-0.098)	0.759 (-0.073-0.055)	0.677 (0.045-0.112)

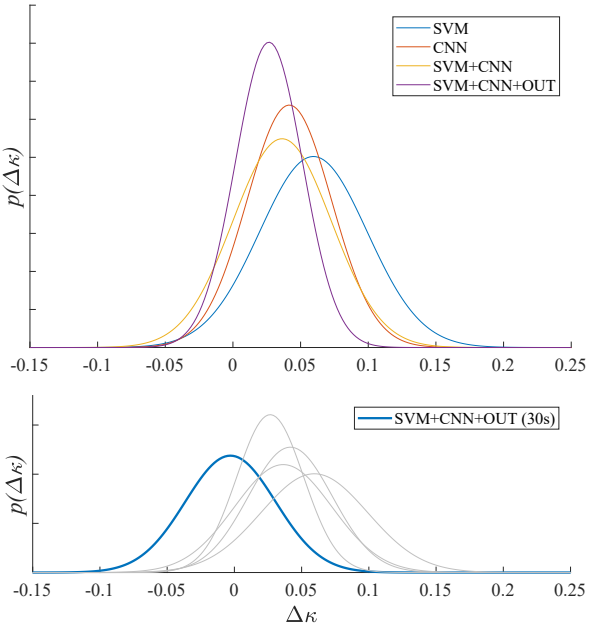

Fig. 2. The performance of the NSDA expressed as differences in inter-observer agreement ($\Delta\kappa$). The upper plot shows $\Delta\kappa$ between human and human/NSDA agreement (NSDA replacing expert 2) for all trialled NSDAs. The lower plot shows the performance increase achieved by increasing the minimum seizure duration from 10s to 30s.

just achieve minimum levels of accuracy with a minimum seizure duration of 10s, the accuracy of the NSDA with a minimum seizure duration of 30s is considerably higher (the 50%CI spans 0) and can confidently be deemed sufficient. There are two reasons why increasing minimum seizure duration (the initial 10s limit was arbitrary); 1) it has been shown that IOA is significantly lower for short duration seizures, and 2) low duration seizures do not meaningfully contribute to seizure burden, the key measure that clinicians aim to minimize with treatment.

We have shown that the output of several NSDAs based on a combination of EEG features and CNN outputs, with or without the inclusion of an outlier detection post-processing stage cannot be statistically differentiated from the annotation of the human expert for neonatal seizure detection. NSDA accuracy is considerably improved when the minimum seizure duration is increased from 10s to 30s. The sufficiency of the NSDA can only be challenged by the ability of the dataset used in this paper to generalize to a larger population of neonatal EEG recordings. The establishment of such an annotated EEG database would be further legitimized with oversight from organizations such as the American Society of Clinical Neurophysiologists or the Task Force on Neonatal Seizures (International League Against Epilepsy) [10], [11].

REFERENCES

[1] R. Wickström, B. Hallberg, M. Bartocci, "Differing attitudes toward phenobarbital use in the neonatal period among neonatologists and child neurologists in Sweden", *Eur J Paediatr Neurol.* vol. 17, pp. 55-63, 2013.

[2] S.R. Mathieson, N.J. Stevenson, E. Low, W.P. Marnane, J.M. Rennie, A. Temko, G. Lightbody, G.B. Boylan, "Validation of an automated seizure detection algorithm for term neonates", *Clin Neurophysiol.* vol. 127, pp. 156-68, 2016.

[3] P. Srinivasakumar, J. Zempel, S. Trivedi, M. Wallendorf, R. Rao, B. Smith, T. Inder, A.M. Mathur, "Treating EEG seizures in hypoxic ischemic encephalopathy: a randomized controlled trial", *Pediatrics*, vol. 136, pp. e1302-9, 2015.

[4] D.M. Murray, G.B. Boylan, I. Ali, C.A. Ryan, B.P. Murphy, S. Connolly, "Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures", *Arch Dis Child - Fetal Neonatal Ed.*, vol. 93, pp. F187-91, 2008.

[5] G.B. Boylan, N.J. Stevenson, S. Vanhatalo, "Monitoring neonatal seizures", *Semin Fetal Neonatal Med*, vol. 18, pp. 202-208, 2013.

[6] A. O'Shea, G. Lightbody, G. Boylan, A. Temko, "Investigating the Impact of CNN Depth on Neonatal Seizure Detection Performance", In *2018 40th Ann Int Conf IEEE Eng Med Biol Soc (EMBC)*, Honolulu, pp. 5862-5865, 2018

[7] K. Tapani, N.J. Stevenson, L. Lauronen, S. Vanhatalo, "Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection". *Int J Neural Sys*, art no. 1850030, 2019.

[8] K. Tapani, N.J. Stevenson, L. Lauronen, S. Vanhatalo, "A dataset of neonatal EEG recordings with seizures annotations", *Sci Data*, vol. 6, art no. 190039, 2019.

[9] N.J. Stevenson, R.R. Clancy, S. Vanhatalo, I. Rosén, J.M. Rennie, G.B. Boylan. "Interobserver agreement for neonatal seizure detection using multichannel EEG", *Ann Clin Transl Neurol* vol. 2, pp. 1002-11, 2015.

[10] R.A. Shellhaas, T. Chang, T. Tsuchida, M.S. Scher, J.J. Riviello, N.S. Abend, S. Nguyen, C.J. Wusthoff, R.R. Clancy. "The American Clinical Neurophysiology Society's guideline on continuous electroencephalography monitoring in neonates", *J Clin Neurophysiol.*, vol. 28, 611-617, 2011.

[11] M.L. Nunes, E.G. Yozawitz, S. Zuberi, E.M. Mizrahi, M.R. Cilio, S.L. Moshé, P. Plouin, S. Vanhatalo, R.M. Pressler, "Neonatal Seizures: Is there a relationship between ictal electro-clinical features and etiology?–A critical appraisal based on a systematic literature review". Epilepsia Open. doi.org/10.1002/epi4.12298, 2019

[12] C.J. Burges. "A tutorial on support vector machines for pattern recognition", *Data mining and knowledge discovery*, vol. 2, pp. 121-167, 1998.