Author/s:
Reuben, C;Karoly, P;Freestone, DR;Temko, A;Barachant, A;Li, F;Titericz, G;Lang, BW;Lavery, D;Roman, K;Broadhead, D;Jones, G;Tang, Q;Ivanenko, I;Panichev, O;Proix, T;Nahlik, M;Grunberg, DB;Grayden, DB;Cook, MJ;Kuhlmann, L

Title:
Ensembling crowdsourced seizure prediction algorithms using long-term human intracranial EEG

Date:
2020-02

Citation:
Reuben, C., Karoly, P., Freestone, D. R., Temko, A., Barachant, A., Li, F., Titericz, G., Lang, B. W., Lavery, D., Roman, K., Broadhead, D., Jones, G., Tang, Q., Ivanenko, I., Panichev, O., Proix, T., Nahlik, M., Grunberg, D. B., Grayden, D. B. ,... Kuhlmann, L. (2020). Ensembling crowdsourced seizure prediction algorithms using long-term human intracranial EEG. Epilepsia, 61 (2), pp.e7-e12. https://doi.org/10.1111/epi.16418.

DR. PHILIPPA J KAROLY (Orcid ID : 0000-0002-9879-5854)

PROF. DAVID B GRAYDEN (Orcid ID : 0000-0002-5497-7234)

DR. LEVIN KUHLMANN (Orcid ID : 0000-0002-5108-6348)

Article type : Brief Communication (includes Case Reports)

Ensembling Crowd-Sourced Seizure Prediction Algorithms Using Long-Term Human Intracranial EEG

Chip Reuben[1], Philippa Karoly[1,2], Dean R. Freestone[1], Andriy Temko[3], Alexandre Barachant[4], Feng Li[5], Gilberto Titericz Jr.[6], Brian W. Lang[7], Daniel Lavery[7], Kelly Roman[7], Derek Broadhead[7], Gareth Jones[8], Qingnan Tang[9], Irina Ivanenko[10], Oleg Panichev[10], Timothée Proix[11,12], Michal Náhlík[13], Daniel B. Grunberg[14], David B. Grayden[2], Mark J. Cook[1], Levin Kuhlmann[1,15,*]

1. Department of Medicine – St. Vincent's Hospital, The University of Melbourne, Parkville VIC 3010, Australia.

2. NeuroEngineering Lab, Department of Biomedical Engineering, The University of Melbourne, Parkville VIC 3010, Australia.

3. Irish Centre for Fetal and Neonatal Translational Research, University College Cork, Cork, Ireland.

4. Grenoble, France.

5. Minnesota, USA.

6. California, USA.

7. Areté Associates, 1550 Crystal Drive, Suite 703, Arlington, VA 22202, USA.

8. UCL Ear Institute, 332 Gray's Inn Road, London WC1X 8EE, UK.

9. Department of Physics, National University of Singapore, Singapore 117551.

10. Kyiv, Ukraine.

11. Department of Neuroscience, Brown University, Providence, Rhode Island, USA.

12. Center for Neurorestoration & Neurotechnology, U.S. Department of Veterans Affairs, Providence, Rhode Island, USA.

13. Prague, Czech Republic.

14. Solverworld, Suite 140, 1337 Mass. Ave, Arlington, Massachusetts, USA.

15. Faculty of Information Technology, Monash University, Clayton VIC 3168, Australia.

* corresponding author

Correspondence to: Levin Kuhlmann, PhD,

Faculty of Information Technology,

Monash University,

Clayton VIC 3183, Australia.

E-mail: levin.kuhlmann@monash.edu

Key Words: epilepsy; seizure prediction; intracranial EEG; refractory epilepsy; Open Data Ecosystem for the Neurosciences; ensemble methods.

Number of text pages: 11

Number of words: 1953

Number of references: 18

Number of figures: 1

Number of tables: 1

Summary

Seizure prediction is feasible but greater accuracy is needed to make seizure prediction clinically viable across a large group of patients. Recent work crowdsourced state-of-the-art prediction algorithms in a worldwide competition, yielding improvements in seizure prediction performance for patients whose seizures were previously found hard to anticipate. The aim of the current analysis was to explore potential performance improvements using an ensemble of the top competition algorithms. The results suggest minor increments in performance may be possible; however, the outcomes of statistical testing limit the confidence in these increments. Our results suggest that for the specific algorithms, evaluation framework and data considered here, incremental improvements are achievable but there may be upper bounds on machine learning-based seizure prediction performance for some patients whose seizures are challenging to predict. Other more tailored approaches that, for example, take into account a deeper understanding of preictal mechanisms, patient-specific sleep-wake rhythms, or novel measurement approaches, may still offer further gains for these types of patients.

Keywords

Epilepsy; seizure prediction; intracranial EEG; refractory epilepsy; Open Data Ecosystem for the Neurosciences; ensemble methods.

Introduction.

Methods for accurate seizure prediction have the potential to transform epilepsy management by offering warnings to patients or triggering interventions.[1-3] Prospective seizure prediction has been shown to be feasible using long-term intracranial electroencephalography (iEEG) recordings obtained during the first-in-human trial of a seizure prediction device, however, prediction accuracy should be increased for a broader range of patients in order to demonstrate clinical utility.[4] Several studies have looked at achieving retrospective improvements with the same long-term data.[5-7] The recent Melbourne-University AES-MathWorks-NIH Seizure Prediction Challenge yielded improvements for the three patients from the NeuroVista trial whose seizures were the hardest to predict. The top

internationally crowd-sourced algorithms increased seizure prediction sensitivity by 90% relative to the original trial, for matched time in warning.[7] However, performance was still not optimal.

Further improvements to prediction accuracy may be made possible by ensembling the top seizure prediction algorithms from this international contest. Ensembling is a general machine learning technique that combines different algorithms in an intelligent manner by merging the complementary outputs of algorithms and/or gain greater confidence in algorithm decisions by weighting the agreement among them.[8,9] Ensembling has yielded boosts in performance in various application domains.[9]

Over 10,000 algorithms were submitted in the aforementioned contest, and eight of the top algorithms were evaluated on held-out data from the same patients.[7] After such extensive development, an important question to answer is whether the performance level obtained from this competition was the upper limit of what is possible for certain patients? To address this question, the current paper aims to determine whether machine learning approaches that ensemble the best performing seizure prediction algorithms from the contest improve on the top performing individual algorithms based on the data from the same three patients who had the lowest seizure prediction performance in the original trial.

Methods.

Data in the form of iEEG were recorded chronically from three female subjects with refractory focal epilepsy from the NeuroVista Seizure Advisory System trial described previously (Table 1a), whose seizures were the hardest to predict.[4,7] These data were used in the 'Melbourne University AES-MathWorks-NIH Seizure Prediction Challenge' hosted on the crowdsourcing platform, Kaggle.com.[7] Contestants were given 10-minute data clips that were labeled either as preictal or interictal in two sets: a training set with the labels and a test set without the labels. In the test set, a randomly chosen subset was used as a private set, for which the contestants had no information about the labels until after the completion of the competition. For each clip, contestants were required to submit the probability that the clip was preictal, referred to as "preictal probability". The final ranking in the competition was determined by the private subset. For the purposes of our analyses, seizure prediction has been defined as the successful identification of a future ictal event within 65 minutes before seizure onset. The criterion for performance was the maximal area under the curve (AUC) of the receiver operating characteristic (ROC) in which thresholding of the range of probabilities yielded a curve of true positive rates versus false positive rates.

After the competition, eight of the top-performing teams participated in an analysis of a much larger, held-out dataset from the same patients. These teams used a variety of machine learning algorithms as previously described.[7] Additionally, circadian weighting was incorporated into a sub-analysis of results in which the original preictal probability predictions from the machine learning algorithms were multiplied by the probability of a seizure occurring at a given time of day.[5,7]

This paper focuses on ensembling the eight algorithms from the aforementioned post-contest held-out data evaluation to investigate improvements in seizure prediction performance. Supervised ensembling[9] was performed with a multilayer perceptron neural network[10] taking as input the set of preictal probabilities for a given ensemble of individual team algorithms for a given data segment. The private contest test set was used as the training set, and the held-out data set was used as the test set. Only private contest test data was used for training of ensemble weighting because this set was the basis of ranking algorithms from the contest. In the neural network, the input layer was equal in size to the number of algorithms in an ensemble, the middle layer size was set to a constant value of 12, and the output layer size was set to 2 for binary classification: preictal or interictal. The objective function (also known as the "cost function") of the neural network was modified such that the errors of the preictal class were multiplied by the ratio of interictal:preictal class in the training set to ensure balanced training data. A sigmoidal activation function was used to obtain the output preictal probabilities. Thus, a single preictal probability was output for each data segment, given an ensemble of input preictal probabilities.

All 247 combinations of at least two team algorithms were explored separately to find which combination achieved the highest performance. For each combination, regularization was used to explore the prospect of improving generalization of the trained models to the test set (i.e. the held-out set), and the regularization parameter λ was set to 0, 0.2, 0.4, 0.6, 0.8, and 1. Each of the corresponding 1482 combinations (247 combinations x 6 λ values) was run in triplicate. All training and testing was performed in a patient-specific manner.

Statistical Methods

A statistical test to compare AUC scores derived from the same data[11-13] was used to assess if the AUC score for the top contest algorithm for a given held-out data set (overall or individual patient) was different from the AUC scores for the ensembled algorithms, both with and without the use of circadian weighting.[5] In addition to AUC-based analysis, clinically relevant pseudo-prospective seizure prediction

performance was evaluated with the metrics of sensitivity (proportion of seizures correctly predicted – i.e. number of seizures occurring during high-seizure-risk divided by the total number of seizures) and proportion of time in warning (i.e. proportion of time in high-seizure-risk warning). Success was defined by an algorithm having higher AUC scores than the original contest algorithms, or higher sensitivities for matching proportion of time in warning. Statistical tests for performance comparisons are reported with a significance level of 0.05 with subsequent Bonferroni correction.[7,13] Pseudo-prospective seizure prediction performance was compared against the performance of random analytic Poisson prediction.[7] See Supplementary Information for performance evaluation details.

Results

In the results without circadian weighting, for patients 1, 2, and 3, and Overall (refers to all data samples independent of patient), 3, 1, 25, and 5 percent of ensembles, respectively, achieved higher AUC scores than the top performing individual team. In the results with circadian weighting, for patients 1, 2, and 3, and Overall, 0.07, 1, 9, and 5 percent of ensembles, respectively, achieved higher AUC scores than the top performing individual team. Overall AUC and individual patient AUC values, with and without circadian weighting, are summarized for the top three ensembles and for the top individual (reference) algorithms in Table 1B. No statistically significant differences were found at the 0.05 significance level after correction for multiple comparisons. Without this correction, however, a total of 58 ensembles of the non-circadian-weighted submissions performed statistically significantly better (ensemble AUCs ranged from 0.77459 to 0.78537 vs. Team A AUC = 0.76151, $p < 0.05$) than did the results of Team A on the overall AUC analyses. A total of 15 ensembles of the circadian-weighted submissions performed statistically significantly better (ensemble AUCs ranged from 0.81267 to 0.81481 vs. Team A AUC = 0.79684, $p < 0.05$) than did the results of circadian-weighted Team A on the overall AUC analyses. Additional analysis provided in the Supplementary Information (Fig S1) revealed no apparent relationships between complementarity of algorithms (captured by average Pearson's correlation of the algorithm predictions) and AUC.

Sensitivity versus proportion of time in warning curves for the top three ensembles, and the top individual team algorithm reference are shown in Fig 1 for the three patients (rows) without (left column) or with (right column) circadian weighting. In all cases, performance was above chance and the best performing ensembles contained the reference team.

Discussion

In the original NeuroVista trial, a machine learning algorithm was used to predict seizures and demonstrate the feasibility of seizure prediction with very high performance for some patients, for example the best case achieved 100% sensitivity and only 3% time in warning.[4] Nevertheless, for patients whose seizures are the hardest to predict, as considered here, more refined approaches are needed to achieve higher performance. Here the approach was ensembling of algorithms.

Although the ensembling methods used in this analysis gave incremental improvements in AUC performance (between 0.5-1.5% depending on the patient) relative to the best individual performing team both with and without circadian weighting, these were not found to be statistically significant when corrected for multiple comparisons. Interestingly, patient 3 had a large number of ensembles with AUC greater than the best original algorithm (25% of ensembles when circadian weighting was not applied). Given the large number of ensembles considered here it is not surprising the statistical results did not survive correction for multiple comparisons; however, the main goal was to consider all possible combinations to search for an upper bound on performance. Our results suggest, for the challenging data considered here, that an upper limit of prediction performance using a black-box, machine-learning approach was reached by the winning competition algorithms.

This does not mean that machine-learning based seizure prediction is not useful, especially given that excellent performance has been achieved with machine learning for some patients as described above. [4] Beyond ensembling, other more refined approaches could be applied to find improvements for these challenging cases. For example, these patients had many seizures that could potentially be characterized and grouped by specific seizure types. Different machine learning algorithms could be trained specifically for each seizure type, such that several algorithms would be used to predict a patient's set of seizure types.[15] It is also very important to note that as a result of Kaggle contest formats, contestants did not have full timing information available to them for training, such as the time of interictal segments relative to preictal segments, which may significantly limit algorithm performance. Moreover, algorithms that directly account for known seizure triggers, such as sleep-wake rhythms, as well as novel measurement approaches may offer alternative routes to performance gains.[2,3,16] In addition, recent work looking at long-term rhythms[18] holds significant promise for seizure forecasting, could be applied in the machine learning context, and suggests that the performance limits seen here may not be tied to potential limits imposed by the physiological processes underlying seizure occurrence.

It is also important to note that a number of the original team algorithms were ensembles themselves. For example, Team A was an ensemble of 11 algorithms from four different individuals. In each of these

cases, the teaming up of individuals was performed towards the end of the competition, so the ensemble algorithms within these individual teams had already been developed independently from one another. Therefore, further performance increments using the ensembling performed in the current study may be limited.

The AUC was not affected by the average correlation of ensemble predictions. This suggests a near-peak level of performance has been obtained by the original algorithms since ensembling complementary algorithms (i.e. those with low average correlation) did not lead to a statistically significant boost in performance. Taken as a whole these results suggest there is likely to be an upper bound on seizure prediction performance for the patients and algorithmic and training approaches considered here. Future work with data from more patients[17] will be needed to truly understand patient-specific or patient-group-specific limits on seizure prediction performance. The online platform Epilepsyecosystem.org[7] and ongoing studies with wearable and implantable devices provide several avenues to push in this direction.

Acknowledgements. This project is supported by the Epilepsy Foundation of America My Seizure Gauge Grant and NHMRC Project Grant GNT1160815. We thank Yuval Kluger and Ariel Jaffe for useful discussions.

Disclosure. D.R.F and M.C. report equity in Seer Medical. All other authors declare no conflicts of interest. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines. The data considered here is accessible online at https://www.epilepsyecosystem.org.

References

1. Mormann, F., Andrzejak, R. G., Elger, C. E., et al. Seizure prediction: the long and winding road. Brain. 2007; 130, 314–333.
2. Freestone DR, Karoly PJ, Peterson ADH, et al. Seizure Prediction: Science Fiction or Soon to Become Reality? Curr Neurol Neurosci Rep. 2015; 15:73
3. Kuhlmann L, Lehnertz K, Richardson M, et al. Seizure prediction — ready for a new era. Nature Reviews Neurology. 2018a; 14: 618–630.
4. Cook MJ, O'Brien TJ, Berkovic SF, et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. Lancet Neurol. 2013; 12: 563-71.

5. Karoly P, Ung H, Grayden DB, et al. The Circadian Profile of Epilepsy Improves Seizure Forecasting. Brain. 2017; 140: 2169-82.
6. Kiral-Kornek I, Roy S, Nurse E, et al. Epileptic seizure prediction using big data and deep learning: toward a mobile system. EBioMedicine. 2018; 27, 103-111.
7. Kuhlmann L, Karoly P, Freestone DR, et al. Epilepsyecosystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG. Brain. 2018b; 141:2619-2630.
8. Dietterich TG. Ensemble methods in machine learning. In International workshop on multiple classifier systems. Springer, Berlin, Heidelberg. 2000; pp 1-15.
9. Zhang C, Yunqian M. eds. Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012.
10. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.
11. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148: 839-43.
12. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LAG, Annane D. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. Critical care medicine. 2013; 41: 954.
13. Brinkmann BH, Wagenaar J, Abbot D, et al. Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. Brain. 2016; 139: 1713-22.
14. Benesty J, Chen J, Huang Y, et al. Pearson correlation coefficient. In Noise reduction in speech processing. Springer, Berlin, Heidelberg, 2009; pp. 1-4.
15. Fisher RS, Cross JH, French JA, et al. Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017; 58, 522-530.
16. Chang WC, Kudlacek J, Hlinka J, et al. Loss of neuronal network resilience precedes seizures and determines the ictogenic nature of interictal synaptic perturbations. Nature Neuroscience. 2018; 21, 1742.
17. Dumanis SB, French JA, Bernard C, et al. Seizure forecasting from idea to reality. Outcomes of the My Seizure Gauge Epilepsy Innovation Institute Workshop. eNeuro. 2017; 4, 0349–0317.
18. Baud MO, Kleen JK, Mirro EA, Andrechak JC, King-Stephans D, Chang EF, et al. Multi-day rhythms modulate seizure risk in epilepsy. Nat Commun 2018; 9: 88.

Table 1. Patient Data and Seizure Prediction Performance.

(A) Patient and seizure data characteristics.

Patient	Age (years)	Gender	Epilepsy Type	Seizures	Lead Seizures	Recording duration (days)	Training clips (% interictal)	Testing clips (% interictal)	Held-out clips (% interictal)
1	22	F	parieto-temporal focal	390	231	559	797 (69.0)	205 (74.1)	12003 (91.2)
2	51	F	occipito-parietal focal	204	186	393	2027(89.2)	994 (94.0)	22630 (96.8)
3	50	F	fronto-temporal	545	216	374	2158 (88.3)	689 (91.3)	25079 (95.6)

(B) Seizure prediction performance on held-out data - ensembles vs originals.

Algorithm	Overall AUC	Algorithm	Patient 1 AUC	Algorithm	Patient 2 AUC	Algorithm	Patient 3 AUC
Best original							
team H	0.76632	Team F	0.77943	Team A	0.82637	Team H	0.7255
Top 3 ensembles							
Team A,F,H; λ=0.4	0.78537	Team D,F,H; λ=1	0.79138	Team A,F,G,H; λ=0.2	0.83075	Team A,B,C,E,H; λ=0.2	0.73477
Team A,B,F,H; λ=0.4	0.78439	Team F,H; λ=1	0.78983	Team A,G; λ=0.4	0.82957	Team A,B,C,D,F,G,H; λ=0.2	0.73415
Team A,F,G,H; λ=0.4	0.78323	Team A,D,F,H; λ=1	0.78674	Team A,G; λ=0.6	0.82951	Team C,D,F,H; λ=0.2	0.73392
Best original with circadian weighting							

team H	0.80208	Team F	0.74009	Team A	0.87458	Team H	0.75977
Top 3 ensembles - circadian weighted							
Team A,F,G,H; λ=0.4	0.81481	Team A,D,F,G,H; λ=0	0.75808	Team A,G; λ=1	0.87943	Team A,F,H; λ=1	0.76354
Team A,F,G,H; λ=0.6	0.81458	Team A,B,D,F,G,H; λ=0	0.75266	Team A,F,G; λ=1	0.87912	Team A,D,F,H; λ=1	0.76339
Team A,F,G,H; λ=0.8	0.81335	Team B,C,F; λ=0	0.74899	Team A,G; λ=0.8	0.87904	Team A,B,C,D,E,H; λ=1	0.76326

Figure Legend

Figure 1. Pseudo-prospective seizure prediction results for the held-out data for the best three ensembles and for the best individual team for each patient without and with circadian weighting. Seizure prediction performances without circadian-weighting for the three patients from the NeuroVista trial whose seizures were the hardest to predict: patients (A) 1 , (C) 2 and (E) 3, and with circadian-weighting for patients (B) 1, (D) 2 and (F) 3. Results are compared to random prediction. The y- and x-axes correspond to sensitivity and proportion of time in warning (i.e. time in high-seizure-risk), respectively. For the different algorithms, data points on the curves correspond to different preictal probability thresholds and only data points surviving correction for multiple comparisons against chance are plotted.