Progress Report
Real-time preictal detection through the application of machine learning to Electroencephalogram signals.

William Riddell
19066041

December 24, 2023

Word Count: 3621

Contents

1 Introduction

Over the last 20 years, Artificial Intelligence (AI) has seen a large evolution through the use of Machine Learning (ML); the statistical analysis of data which leads to the unveiling of characteristics and connections. (Awad & Khanna 2015). There has been a large uptake of applying ML techniques to biomedical data, increasing the speed and accuracy of prediction, detection, diagnosis, and prognosis.

Electroencephalograms (EEGs) measure the electrical signals in the brain. EEGs have a great use in giving an insight into the inner workings of the brain, for example allowing us to pick up abnormalities preceding and during their occurrence. "A seizure is a burst of uncontrolled electrical activity between brain cells (also called neurons or nerve cells) that causes temporary abnormalities in muscle tone or movements (stiffness, twitching or limpness), behaviours, sensations or states of awareness." (Medicine n.d.) Due to this, monitoring the brain's electrical activity through the use of an EEG, and applying analysis through an ML model may allow us to detect the preictal period. "An automated accurate prediction of seizures will significantly improve the quality of life of patients and reduce the burden on caregivers" (Acharya, Hagiwara & Adeli 2018)

1.1 Background

"Because of their unpredictable nature, uncontrolled seizures represent a major personal handicap and source of worry for patients. In addition, persistent seizures constitute a considerable burden on healthcare resources." (Assi, Nguyen, Rihana & Sawan 2017) Due to this both medication and surgery are available to applicable patients, although with 30% patients being refractory to drug therapy, and an equally bleak surgery success rate; 75% in lesional cases, and 50% in nonlesional cases for temporal lobe cases along with 60% in lesional cases and merely 35% in nonlesional for frontal lobse cases (Assi et al. 2017), a large population of patients would greatly benefit from the prediction of their uncontrollable seizures, along with an relief of burden for the healthcare system when working with seizure patients.

1.2 Aim and Objectives

This project will aim to develop an consistent ML model which triggers an alert if a preictal period is detected. The model will have to achieve a high degree of accuracy ($\geq 90\%$) when being applied to real-time EEG data. To aid in achieving this an ML model comparison will be realized, in which an selection of ML models will be developed and applied to a chosen dataset, and their results evaluated.

Objectives

1. Create a preprocessing pipeline which extracts the dataset into an ML interpretable format such as a CSV, as well as appending preictal classification labels to the extracted data.

2. Develop a feature extraction process which works in real time. The features which are extracted should have lead to success for other researchers.

3. Conduct an experiment where ML models are applied to the extracted features, and compare the results. The highest performing model is selected to be further developed.

4. Tune the hyper-parameters of the selected ML model.

5. Realize and implement the extensions to enable the ML model to work in a real time setting.

1.3 Project Requirements

The final product needs to have the ability to accept a stream of raw EEG data in an CSV format. From the stream of data it will have to extract statistical features within the frequency of the stream and use them, along side the raw data to classify the current state of the patient. If a preictal state is detected through classification an alert is displayed. This has to happen within 5 seconds of a preictal state starting. The final model is required to have a detection success rate of $\geq 90\%$ on testing data from the same dataset.

2 Background Review

2.1 Datasets

(Wong, Simmons, Rivera-Villicana, Barnett, Sivathamboo, Perucca, Ge, Kwan, Kuhlmann, Vasa et al. 2023) reviews 10 datasets available to download. It evaluates the way the EEGs were physically setup on the subject, the subjects themselves and the data's properties. Wong et al. also states their opinion on what tasks suit what dataset, with the main two tasks being either detection or prediction.

Dataset
University of Bonn
CHB-MIT Scalp EEG
Melbourne-NeuroVista seizure trial (Neurovista Ictal)
Kaggle UPenn and Mayo Clinic's Seizure Detection Challenge
Neurology and Sleep Centre Hauz Khas
Kaggle American Epilepsy Society Seizure Prediction Challenge
Kaggle Melbourne-University AES-MathWorks-NIH Seizure Prediction Challenge
TUH EEG Seizure Corpus (TUSZ)
Siena Scalp EEG
Helsinki University Hospital EEG

Table 1: The Datasets analysed

Within these datasets Wong et al. was also able to find the way the EEG nodes were positioned on the subject's cranium, along with whether the EEG nodes were either placed intracranial or extracranial. Wong et al. also the number of channels that are contained in the raw EEG data for each dataset.

Dataset	**Number of channels**	**Placement method**	**Type of signal**
University of Bonn	1	International 10–20 system, Intracranial	Scalp/Intracranial EEG
CHB-MIT Scalp EEG	18	International 10–20 system/Nomenclature	Scalp EEG
Melbourne-NeuroVista seizure trial (NeuroVista Ictal)	16	Intracranial	Intracranial EEG
Kaggle UPenn and Mayo Clinic's Seizure Detection Challenge	16–76	Intracranial	Intracranial EEG
Kaggle American Epilepsy Society Seizure Prediction Challenge	16	Intracranial	Intracranial EEG
Neurology and Sleep Centre Hauz Khas	1	International 10–20 System	Scalp EEG
Kaggle Melbourne-University AES-MathWorks-NIH Seizure Prediction Challenge Data	16	Intracranial	Intracranial EEG
TUH EEG Seizure Corpus (TUSZ)	23–31	International 10–20 system / Nomenclature	Scalp EEG
Helsinki University Hospital EEG	19	International 10–20 system	Scalp EEG
Siena Scalp EEG	20/29	International 10–20 system/Nomenclature	Scalp EEG

Table 2: Channel Characteristics

Wong et al. also noted along with this data that the "University of Bonn dataset contains a mixture of both scalp and intracranial EEG data where scalp EEG from healthy subjects was taken, while intracranial EEG was taken from subjects with a history of seizures." (Wong et al. 2023). This

may present a skew on the ML model during training.

Dataset	**Noncontinuous data**	**Short-term continuous data**	**Continuous data**
University of Bonn	Yes	No	No
CHB-MIT Scalp EEG	No	Yes	Yes
Melbourne-NeuroVista seizure trial (Neurovista Ictal)	N/A	N/A	N/A
Kaggle UPenn and Mayo Clinic's Seizure Detection Challenge	Yes	No	No
Kaggle American Epilepsy Society Seizure Prediction Challenge	Yes	No	No
Neurology and Sleep Centre Hauz Khas	Yes	No	No
Kaggle Melbourne-University AES-MathWorks-NIH Seizure Prediction Challenge Data	Yes	No	No
TUH EEG Seizure Corpus (TUSZ)	No	Yes	No
Helsinki University Hospital EEG	No	Yes	No
Siena Scalp EEG	No	Yes	No

Table 3: Temporal properties

Wong et al. ordered the datasets into groups, either continuous or non continuous data. For the continuous data they separated out datasets which record for less that 24 hours in a single go, these were classified as "Short-term continuous" data.

Dataset	**Number of subjects**	**Subject type**
University of Bonn	10	Human
CHB-MIT Scalp EEG	23	Human
Melbourne-NeuroVista seizure trial (NeuroVista Ictal)	12	Human
Kaggle UPenn and Mayo Clinic's Seizure Detection Challenge	12	Human & Canine
Kaggle American Epilepsy Society Seizure Prediction Challenge	7	Human & Canine
Neurology and Sleep Centre Hauz Khas	10	Human
Kaggle Melbourne-University AES-MathWorks-NIH Seizure Prediction Challenge Data	3	Human
TUH EEG Seizure Corpus (TUSZ)	642	Human
Helsinki University Hospital EEG	79	Human
Siena Scalp EEG	14	Human

Table 4: Subject properties

Wong et al. also was able to identify the number of subjects within each dataset. Within the two "Kaggle" datasets there are Canine subjects, making them unsuitable for this project.

Within the review, they also produced tables displaying the segment information for each dataset, breaking down the recording length and frequency, along with the number of events and segments. This information should not weight into which dataset suits the idea of preictal prediction so shall be left out in this background review. Wong et al. also discussed the idea of the class imbalance problem, where the number and length of each ictal period is unbalanced. Two datasets, "University of Bonn" and the "Neurology and Sleep Centre Hauz Khas" have addressed this issue and have balanced their data between ictal, preictal, interictal and nonictal periods.

Taking the research into account Wong et al. suggested which dataset

suits either prediction or detection. "Since the aim of seizure prediction is to forecast impending seizures, EEG recordings that include preictal and interictal data should be used for the study, while the aim of seizure detection is to detect ongoing seizure events, hence, EEG recordings that contain ictal and interictal data should be used." (Wong et al. 2023).

Dataset	**Application**
University of Bonn	Seizure detection
CHB-MIT Scalp EEG	Seizure detection/Prediction
Melbourne-NeuroVista seizure trial (NeuroVista Ictal)	Seizure detection/Prediction
Kaggle UPenn and Mayo Clinic's Seizure Detection Challenge	Seizure detection
Kaggle American Epilepsy Society Seizure Prediction Challenge	Seizure prediction
Neurology and Sleep Centre Hauz Khas	Seizure detection/Prediction
Kaggle Melbourne-University AES-MathWorks-NIH Seizure Prediction Challenge Data	Seizure prediction
TUH EEG Seizure Corpus (TUSZ)	Seizure detection/Prediction
Helsinki University Hospital EEG	Seizure detection/Prediction
Siena Scalp EEG	Seizure detection/Predictio

Table 5: Suggested applications

2.2 Machine Learning (ML) Models

A series of papers have been reviewed with the following ML model types being used:

- K Nearest Neighbor (kNN)
- Support Vector Machine (SVM)
- Logistical Regression (LR)
- Random Forest (RF)
- Artificial Neural Network (ANN)

- Convolutional Neural Network (CNN)

K Nearest Neighbor (kNN)

(Wang, Chaovalitwongse & Wong 2013) shows an kNN classifier being used which lead to a sensitivity of 73% and a specificity of 67% when using the estimation of short term maximum lyapunov exponent.

Support Vector Machine (SVM)

Various SVM classifiers were found to be used, one of which was by (Cho, Min, Kim & Lee 2016) which used both clinical and generated data, where the training data was filtered using BPF, EMD, MEMD and NA-MEMD. Features were then extracted through EMD, MEMD, and frequency selection.(Cho et al. 2016) was able to achieve a classification rate of around 80%.

Logistical Regression (LR)

Mirowski et al. used a combination of LR and SVM classifiers to achieve a 71% accuracy along with 0 false positives (Mirowski, Madhavan, LeCun & Kuzniecky 2009).

Random Forest (RF)

Jacobs et al. used a RF derived approach which was a multistage classification system that used cross-frequency coupling. Jacobs et al. was able to achieve an impressive 87.9% sensitivity, and an 93.4% area-under-the-ROC (Jacobs, Hilton, Del Campo, Carlen & Bardakjian 2018).

Artificial Neural Network (ANN)

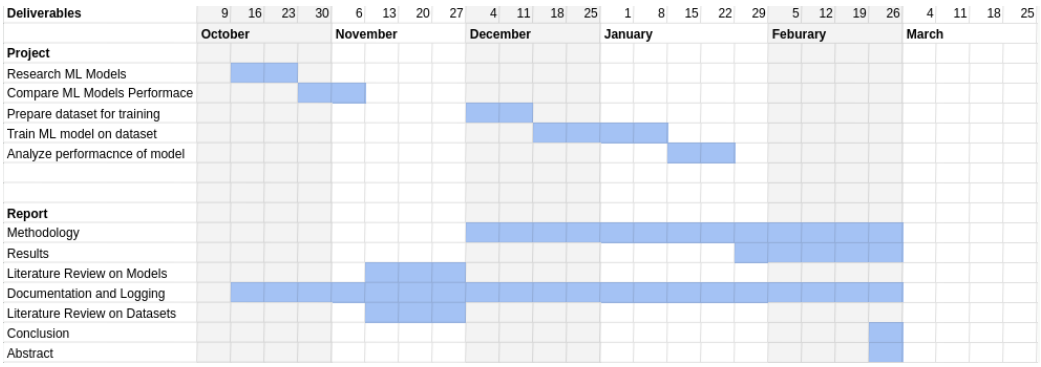
ANN have enabled researchers to develop increasingly accurate models. (**?**) shows this in their results boasting an accuracy of 92.3% and an sensitivity of 100%. Sharma however did use 72 parameters for classification.

Convolutional Neural Network (CNN)

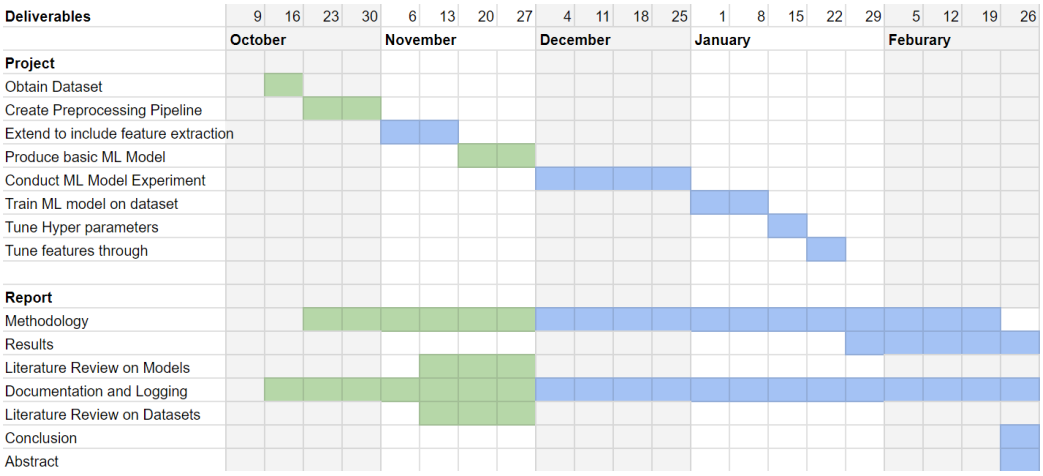
Various CNN have been developed, with most of them showing promising results. An example of this was from (Mirowski et al. 2009), using the same dataset Mirowski et al. used to train their RF based model off, they were able to achieve a 100% prediction rate for 15 out of their 21 patients, along with 0 false positives.

It should be noted, that Mirowski et al. training this model for a per patient basic, and suggested that each patient has their own compilation of models trained to them for the highest chance of prediction, along with minimal false positives.

2.3 Schedule



The original Gantt chart above has been modified to more closely reflect the changes in schedule which have been realized as the projects developed. The new Gantt chart has attempted to keep a time frame consistent with the original.



There has been decent advances for the project; A system has been developed which extracts raw data from the "CHB-MIT Scalp EEG" Dataset and appends classification labelling onto the CSVs which contain the extracted data. The labelled dataset was then used to train a SVM. The SVM was able to achieve surprisingly high results from a small subset of 3 hours, although due to the small dataset over fitting was almost definitely occurring. Moving forward into the new year the project will need to see feature extraction implemented and applied to the entire labeled raw data. Once features have been extracted another SVM can be trained which will produce a more representable result of the final model. Furthermore once the features have been extracted ML model experiment will be able to be conducted.

As the "Report" section of the Gantt chart requires, basic documentation has been created that shows how to install dependencies and run the system, along with what each modifier in the configuration file changes. A log of my activities has also been created which shows the methodology of the process on a monthly time frame.

2.4 Future Activities

1. Feature Extraction:
    - Evaluate other papers to decide on which features to extract.
    - Build an feature extraction pipeline.

2. Model Development:

    - Train an SVM model against the feature extracted dataset.
    - Create an experiment to derive the most successful ML model.
    - Train and tune the ML model's hyper parameters.

2.5 Problems Encountered

When attempting to gain access to the prediction applicable datasets that were discussed in the literature review, some datasets were not public; Contacting the managing bodies of a few of these datasets did not lead to a response, translating to an reduction of available data to train against. This was a factor when deciding to use the "CHB-MIT Scalp EEG" Dataset. This dataset however did not come without its issues.

"CHB-MIT Scalp EEG" came with a descriptive labelling of each EEG data file. The description of each file included the ictal periods start times, the number of ictal periods, the channel names, and the frequency of the recorded data. Some of the descriptions were not accurate of the raw data it was attempting to describe. An example of this was that the descriptor file contained duplicate channel names. This caused issues when extracting the raw data as the Python 3 library "mne" requires unique channel names, this was easily solved through an implimentation of a naming convention in this case. Another issue which arose was the number of raw channels extracted was greater than the expected number of channels described in the description file. This again caused issues with "mne". In these cases however the raw extracted channels contained a basic name that could be used, although for other datasets where this may be an issue another naming convention may have to be implemented.

Feature extraction has not been fully implemented yet, one of the concerns of feature extraction however will be the speed of the extraction in relation to time, particularly, if the extraction of the statistical features can be done within the frequency of the EEG recording. As the "CHB-MIT Scalp EEG" dataset has a frequency of 256hz, feature extraction and also the writing of said features all has to occur within 3.9ms. Due to this a solution will have to be realized. The current approach is to develop a Python3 program to achieve this and benchmark the speed, the expectation is that the Python3 program will be drastically too slow, but gives an indication to

if feature extraction can be achieved for each frequency through possibly a C program. If not, other avenues will have to be explored, such as creating hash maps for faster approximated feature extraction, or even sampling a smaller number of recorded rows. Principal Component Analysis (PCA) may have to be explored here, as extracting fewer features, but for a greater number of frequencies may lead to higher accuracy results, although tuning of features will have to take place after the ML model experiment.

2.6 Development System Progress

A spiral development system has been adopted for this project, with there being a small number of large iterations. The inital cycle has been complete, as a basic ML model has been developed and trained against the preprocessed data. For the next itteration feature extraction needs to occour, and another ML model will need to be trained. Another cycle will contain the ML model experiment where different models' performances are compared, allowing for the thirf iteration to begin. The more accurate ML model should then be fully implemented. Other iterations include PCA of features, and hyper parameter tuning.

2.7 Supervision

The project supervisor has aided the direction of the project, introducing the idea of setting up the ML model experiment to evaluate the accuracy of each model. This has lead to further realizations such as PCA experiments which will help solve upcoming issues discussed in the "Problems Encountered" section. The project supervisor also gave literature review feedback which pushed me to analyse currently developed systems I would not of otherwise. Weekly meetings with the supervisor kept them informed on the progress of the project, allowing them to give guidance as discussed.

3 Professional Issues and Risks

3.1 Professional Issues

1. Legal Issues:

- There are a few legal issue which are attached to the project. Firstly GDPR rules need to be followed as the data moving into the model for both live classification and training need to be dealt in a legal manner. Data will be stored after classification as this allows for further development of the model, therefore an appropriate timespan will need to be decided before the data needs to be deleted. Along with this, security measures should be implemented to to stop attackers or un-authorized persons viewing or copying the data. The other legal issue which may come into play would be who is liable if a seizure is avoided. This could be solved through an agreement the end user has to accept stating the producers of the product, including the developers as not liable if this occurs as the product will never be truly %100 accurate.

2. Social Issues:

   - People from all walks of life have epilepsy, and due to financial costs some may not be able to afford the final solution. This will be an issue that may not be solvable by the producers of the software, and instead may need to rely on outsourced funding such as the healthcare service. The producers can make the software available, although it will still take time, money, and knowledge for the individual to setup a system that meets the expectations of a finished product.

3. Ethical Issues:

   - The final product should attempt to achieve the same accuracy regardless of age or gender, although due to the differences in the the human brain this may be a difficult task to achieve for a single solution. In an ideal world a model will be trained, focused on different populations to pick up on their differencing characteristics of their preictal periods, although due to the lack of datasets this currently is not achievable until more data is recorded once the initial product has been deployed.

4. Environmental Issues:

   - EEG nodes are comparatively inexpensive to their possible benefits, meaning the creation of the device won't have a large overall

cost. The ML model training however will be computationally expensive, translating to large energy usage, therefore ways to minimize energy consumption when training the ML model should be taken into account for the final product.

5. Intellectual Property Issues:

   - There will be an issue when deciding who owns the recorded data. If the company who produces the final product owns it then they can utilize the large amount of data to further refine and train more advanced models, allowing them to tackle other issues such as the ones discussed in "Social Issues", although this should be a choice for the end user. Another issue may be patenting the final product and which components should be patented or copyrighted such if the most recent model or older versions will be freely available, or if they will be closed sourced.

6. Accessibility Issues:

   - As discussed epileptic patients come from all walks of life which means the final device needs to have an accessible interface, allowing everyone to have a clear indication of the state of the device, including the preictal period alert or even if the device is on. Different people will need different alert methods, and the final device should be extensible, allowing the final patient to fit it to their disability. Some variants of the device should have any combination of light, sound and vibration alert, as well as notifications to any of their devices.

3.2 Risks

3.2.1 Risk Matrix

The numbers in each cell of the risk matrix corresponds to the items in the risk assessment.

Probability	Harm Severity			
	Minor	Marginal	Critical	Catastrophic
Certain				
Likely				
Possible				
Unlikely			4, 6	3, 5, 7
Rare				1, 2

Table 6: Risk Matrix

3.2.2 Risk Assessment

The risks proposed are for a final product which is available to consumers who suffer from epilepsy as well as healthcare services. The current scope of the project does not affect the risks stated above, and therefore the development goals have not been changed to negate any of these risks.

No.	Risk	Impact	Mitigation Strategy
1	Personal data could be leaked through an attack	Fail to adhere to GDPR. Criminal Offence	Implement encryption for data as well as increasing security measures
2	Personal data is leaked by unauthorized employee	Fail to adhere to GDPR. Criminal Offence	Tighten access controls. Educate employees about password and general security
3	Personal data is leaked by an authorized employee	Fail to adhere to GDPR. Criminal Offence	Employ education techniques stating the importance of adhering to GDPR as well as minimizing access to sensitive data.
4	False Negatives for the individual	End users will be unprepared for their seizures and may be caught in an unfavourable situation depending on reliance on the final product.	Implement cross validation techniques. Request all missed seizures to be logged or automatically detected for further training and inspection.
5	False Positives for the individual	A patient may experience un-needed stress or stop important activities due to false positives. Could have measurable knock on effects for an individual.	False positives should be recorded which will allow for further development of the model.
6	False Negatives in a healthcare environment	Staff may not be prepared for an seizure, increasing reaction times	See "False Negatives for the individual"
7	False Positives in a healthcare environment	Staff may waste time preparing for a seizure which never occours, where their help may be needed elsewhere	See "False Positives for the individual"

Table 7: Risk Assessment

Acronyms

AI Artificial Intelligence. 3

ANN Artificial Neural Network. 9, 10

CNN Convolutional Neural Network. 10, 11

EEG Electroencephalogram. 3, 4

EEGs Electroencephalograms. 3, 5

EMD Empirical Mode Decomposition. 10

kNN K Nearest Neighbor. 9, 10

LR Logistical Regression. 9, 10

MEMD Multivariate Empirical Mode Decomposition. 10

ML Machine Learning. 2–4, 7, 9, 12–14, 16

PCA Principal Component Analysis. 14

RF Random Forest. 9–11

SVM Support Vector Machine. 9, 10, 12, 13

References

Acharya, U. R., Hagiwara, Y. & Adeli, H. (2018), 'Automated seizure prediction', *Epilepsy & Behavior* **88**, 251–261.

Assi, E. B., Nguyen, D. K., Rihana, S. & Sawan, M. (2017), 'Towards accurate prediction of epileptic seizures: A review', *Biomedical Signal Processing and Control* **34**, 144–157.

Awad, M. & Khanna, R. (2015), *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, Apress.
URL: *https://books.google.co.uk/books?id=qPGQnAEACAAJ*

Cho, D., Min, B., Kim, J. & Lee, B. (2016), 'Eeg-based prediction of epileptic seizures using phase synchronization elicited from noise-assisted multivariate empirical mode decomposition', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **25**(8), 1309–1318.

Jacobs, D., Hilton, T., Del Campo, M., Carlen, P. L. & Bardakjian, B. L. (2018), 'Classification of pre-clinical seizure states using scalp eeg cross-frequency coupling features', *IEEE Transactions on Biomedical Engineering* **65**(11), 2440–2449.

Medicine, J. H. (n.d.), 'Types of seizures, what is a seizure?'.
URL: *https://www.hopkinsmedicine.org/health/conditions-and-diseases/epilepsy/types-of-seizures*

Mirowski, P., Madhavan, D., LeCun, Y. & Kuzniecky, R. (2009), 'Classification of patterns of eeg synchronization for seizure prediction', *Clinical neurophysiology* **120**(11), 1927–1940.

Wang, S., Chaovalitwongse, W. A. & Wong, S. (2013), 'Online seizure prediction using an adaptive learning approach', *IEEE transactions on knowledge and data engineering* **25**(12), 2854–2866.

Wong, S., Simmons, A., Rivera-Villicana, J., Barnett, S., Sivathamboo, S., Perucca, P., Ge, Z., Kwan, P., Kuhlmann, L., Vasa, R. et al. (2023), 'Eeg datasets for seizure detection and prediction—a review', *Epilepsia Open* .