Project Proposal
Real-time preictal detection through the application of machine learning to Electroencephalogram signals.

William Riddell

October 17, 2023

Comments for Kashi will be written in italics. - William

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)

This project will aim to develop an 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. Throughout this project I will compare previous attempts using different ML models, and I will evaluate the different datsets avaliable for preictal prediction.

2 Background Review

(Wong, Simmons, Rivera-Villicana, Barnett, Sivathamboo, Perucca, Ge, Kwan, Kuhlmann, Vasa et al. 2023) reviews 10 datasets avaliable 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 suite what dataset, with the main two tasks being either detection and 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: 10 Datasets within (Wong et al. 2023)

Within these datasets Wong et al. was also able to find the way the EEG nodes were posistioned 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: (Wong et al. 2023)

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: (Wong et al. 2023)

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

Dataset CHB-MIT (Shoeb 2009) (Goldberger, Amaral, Glass, Hausdorff, Ivanov, Mark, Mietus, Moody, Peng & Stanley 2000 (June 13)) was taken from Boston Children's Hospital. It is a long-term dataset with recordings of 22 paediatric subjects, 5 male and 17 female, who have intractable seizures. The dataset contains 23 EEG signals positioned on the scalp in the international 10-20 EEG system (Sharbrough 1991). The dataset is labelled with the timestamps of the onset and ending of each seizure. Over the course of several days the onsets and ends of a total of 182 seizures were annotated onto the long-term EEG data (Shoeb 2009) (Goldberger et al. 2000 (June 13)). This dataset is utilized by many papers when working on real-time epileptic seizure detection, there have been many different approaches with detection accuracy reaching 92.23% (Usman, Usman, Fong et al. 2017)

Usman et al. was able to achieve a preictal detection rate (sensitivity) of 92.23% against the CHB-MIT dataset. (Usman et al. 2017) Usman noted that poor Signal-to-Noise Ratio (SNR) within the dataset lead to a large amount of false positives and posed as a issue during his research. To combat this he combined the many EEG signals into a single surrogate signal and then utilized Empirical Mode Decomposition (EMD) to increase the SNR. Usman et al. then broke the surrogate signal down and "extracted multiple features including entropy, approximate entropy, and Hjorth parameters." In Usman's et al. papers, it was observed that both statistical and spectral features played a part in detecting the differences between interictal and preictal states so he also extracted spectral moments, and statistical moments to aid with his classification. He finally utilized a Support Vector Machine (SVM) taking in a continuous EEG signal broken into non-overlapping one second sections and classifying if it was part of a preictal period (Usman et al. 2017).

Gao at al. also used the CHB-MIT dataset to train various ML models, although they developed a system which assigns weights to each sample in the dataset, and then uses a Genetic Algorithm (GA) to optimize the weights according to the increase of accuracy of prediction when applied to their validation set (Gao, Liu, Cui, Qian & Chen 2022). Gao's et al. approach solved some common issues that surface when attempting to predict preictal periods; EEG signals when monitoring epileptic seizures face an class imbalance problem, to address this researchers normally reduce the number of interictal samples (Ozcan & Erturk 2019), or increase the number of preictal periods, often using generative methods (Usman, Khalid & Bashir 2021), but as Gao et al. stated, this increased the uncertainty in the model. Goa et al. approach of assigning weights to samples avoids this downfall. Another benefit of applying Gao at al. methodology is the avoidance of the noisy label problem. This is where there is no standardized preictal period, with most researchers applying their own empirical fixed time length before the seizure. This can vary from 2 minutes before the ictal period to 30 minutes. By applying weights to different samples the ML model "can select and focus on more valuable samples for seizure prediction. These samples prompt the model to perform better on the validation set, thus further improving the generalization of the model." (Gao et al. 2022). Gao et al. was able to achieve a sensitivity of 85.9%, and a false positive rate / hour of 0.09.

Other methods for predicting seizures exist, although they are outside the realm of ML. When a seizure occurs, or during the preictal state, the spike rate and variation in the EEG signals change (Lange, Lieb, Engel Jr & Crandall 1983) (Truccolo, Donoghue, Hochberg, Eskandar, Madsen, Anderson, Brown, Halgren & Cash 2011) "spike rate is used as the indicator to anticipate seizures in Electroencephalogram (EEG) signal. Spikes detection step is used in EEG signal during interictal, preictal, and ictal periods followed by a mean filter to smooth the spike number. The maximum spike rate in interictal periods is used as an indicator to predict seizures." (Slimen, Boubchir & Seddik 2020). Using this method, Slimen et al. was able to achieve a 92% accuracy. *This is an interesting approach, although not ML, I wonder if I can use it to aid in my ML methodology?*

3 Methodology

3.1 Approach (Description of the research and development methodology, e.g. Software development model, requirement gathering method, test and evaluation process)

I will be utilizing the open-source CHB-MIT dataset (Shoeb 2009) (Goldberger et al. 2000 (June 13)) as it contains all ictal periods which is ideal for the prediction of epileptic seizures. I will be following the preprocessing stages that Usman et al. and Gao et al used. This involves converting the EEG signals into a single signal through the use of applying an averaging filter and the Common Spatial Pattern (CSP) algorithm to increase the SNR. Once I have applied these algorithms I will perform EMD which breaks the surrogate signal into its oscillatory functions, which are amplitude, period, and frequency. This is also known as its Intrinsic Mode Functions (IMF). I will also be applying the sample weighting methodology used in (Gao et al. 2022) to aid in my prediction.

	Features
Statistical moment	Mean
	Variance
	Skewness
	Kurtosis
Spectral band power	Delta (0.5 – 4 Hz)
	Theta (4–8 Hz)
	Alpha (8–13 Hz)
	Beta (13–30 Hz)
	Gamma-1 (30–50 Hz)
	Gamma-2 (50–75 Hz)
	Gamma-3 (75–100 Hz)
	Gamma-4 (100–128 Hz)
Hjorth parameters	Mobility
	Complexity

Table 4: Proposed Features to extract

I will be programming the preprocessing and the ML model in Python 3 (Van Rossum & Drake 2009), using library's such as TensorFlow (Abadi, Agarwal & et al. 2015) which I will be using to construct my SVM. I will also be using Numpy (Harris, Millman & van der Walt et al. 2020) to write in my averaging filter and CSP algorithm. I will also be using the "EMD-signal" library (Laszuk 2017) to split the surrogate signal into its IMF. Using the Leaving One Out Cross-Validation (LOOCV) methodology I will be fine tuning the ML model against different sample weights as stated in (Gao et al. 2022).

I will be version managing this project using the software "Git" , and hosting the repository on "Github", and I will be using my own hardware, a Nvidia GeForce GTX 1080, to train the SVM. For testing purposes I will be using the libraray PyTest to unit test my functions, and as mentioned above I will be using LOOCV to test my ML model.

4 Project management

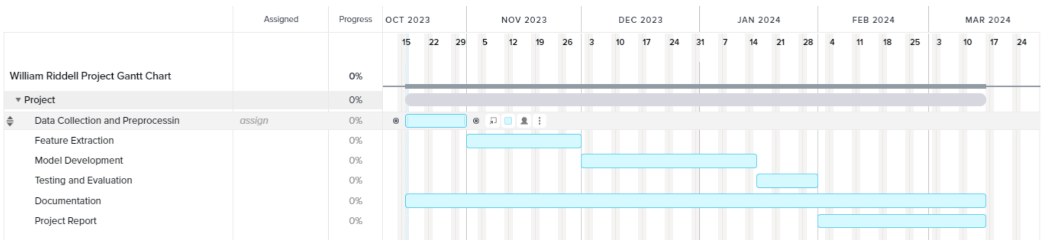
4.1 Activities

1. Data Collection and Preprocessing:

- Download the CHB-MIT dataset.
- Preprocess EEG signals:
  - Apply averaging filter to combine EEG signals.
  - Implement Common Spatial Pattern (CSP) algorithm for noise reduction.
  - Perform Empirical Mode Decomposition (EMD) to extract Intrinsic Mode Functions (IMF).

2. Feature Extraction:
   - Extract features. See Table 1.
   - Apply the sample weighting methodology used in (Gao et al. 2022) for feature weighting.
3. Model Development:
   - Train a Support Vector Machine (SVM) using the preprocessed features.
   - Tune SVM hyper-parameters for optimal performance using LOOCV.
4. Testing and Evaluation:
   - Evaluate the SVM model using performance metrics.
   - Validate the prediction accuracy and sensitivity with a test dataset.
5. Documentation:
   - Maintain project logs, reports, and literature references within the Git repository.
   - Document the entire process throughout the development.
6. Project Reporting:
   - Create a final project report summarizing the methodology, results, and conclusions.
   - Implement any additional information, such as documentation and sources.

4.2 Schedule



4.3 Deliverables

- A well-documented real-time epileptic seizure prediction model using EEG signals from the CHB-MIT dataset.
- A final project report outlining the methodology, results, and conclusions.
- Included in the project report, logs documenting activities, tasks, and changes throughout the project.
- Organized data, including raw data, the preprocessed data, and relevant literature.

Acronyms

AI Artificial Intelligence. 1

CSP Common Spatial Pattern. 6, 7

EEG Electroencephalogram. 1, 2, 4–6

EEGs Electroencephalograms. 1

EMD Empirical Mode Decomposition. 5, 6

GA Genetic Algorithm. 5

IMF Intrinsic Mode Functions. 6, 7

LOOCV Leaving One Out Cross-Validation. 7, 8

ML Machine Learning. 1, 2, 4–7

SNR Signal-to-Noise Ratio. 5, 6

SVM Support Vector Machine. 5, 7

References

Abadi, M., Agarwal, A. & et al., P. B. (2015), 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: *https://www.tensorflow.org/*

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

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*

Chakraborti, S., Choudhary, A., Singh, A., Kumar, R. & Swetapadma, A. (2018), 'A machine learning based method to detect epilepsy', *International Journal of Information Technology* **10**, 257–263.

Gao, Y., Liu, A., Cui, X., Qian, R. & Chen, X. (2022), 'A general sample-weighted framework for epileptic seizure prediction', *Computers in Biology and Medicine* **150**, 106169.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K. & Stanley, H. E. (2000 (June 13)), 'PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals', *Circulation* **101**(23), e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Gupta, V. & Pachori, R. B. (2019), 'Epileptic seizure identification using entropy of fbse based eeg rhythms', *Biomedical Signal Processing and Control* **53**, 101569.

Harris, C. R., Millman, K. J. & van der Walt et al., S. J. (2020), 'Array programming with NumPy', *Nature* **585**(7825), 357–362.
URL: *https://doi.org/10.1038/s41586-020-2649-2*

Kumar, A. & Kolekar, M. H. (2014), Machine learning approach for epileptic seizure detection using wavelet analysis of eeg signals, *in* '2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)', IEEE, pp. 412–416.

Lange, H. H., Lieb, J. P., Engel Jr, J. & Crandall, P. H. (1983), 'Temporo-spatial patterns of pre-ictal spike activity in human temporal lobe epilepsy', *Electroencephalography and clinical neurophysiology* **56**(6), 543–555.

Laszuk, D. (2017), 'Python implementation of empirical mode decomposition algorithm', https://github.com/laszukdawid/PyEMD.

Li, C., Zhou, W., Liu, G., Zhang, Y., Geng, M., Liu, Z., Wang, S. & Shang, W. (2021), 'Seizure onset detection using empirical mode decomposition and common spatial pattern', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **29**, 458–467.

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*

Ozcan, A. R. & Erturk, S. (2019), 'Seizure prediction in scalp eeg using 3d convolutional neural networks with an image-based approach', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27**(11), 2284–2293.

Samiee, K., Kiranyaz, S., Gabbouj, M. & Saramäki, T. (2015), 'Long-term epileptic eeg classification via 2d mapping and textural features', *Expert Systems with Applications* **42**(20), 7175–7185.

Sharbrough, F. (1991), 'American electroencephalographic society guidelines for standard electrode position nomenclature', *Clin Neurophysiol* **8**, 200–202.

Shen, M., Wen, P., Song, B. & Li, Y. (2022), 'An eeg based real-time epilepsy seizure detection approach using discrete wavelet transform and machine learning methods', *Biomedical Signal Processing and Control* **77**, 103820.

Shoeb, A. H. (2009), Application of machine learning to epileptic seizure onset detection and treatment, PhD thesis, Massachusetts Institute of Technology.

Shoeb, A. H. & Guttag, J. V. (2010), Application of machine learning to epileptic seizure detection, *in* 'Proceedings of the 27th international conference on machine learning (ICML-10)', pp. 975–982.

Slimen, I. B., Boubchir, L. & Seddik, H. (2020), 'Epileptic seizure prediction based on eeg spikes detection of ictal-preictal states', *Journal of biomedical research* **34**(3), 162.

Truccolo, W., Donoghue, J. A., Hochberg, L. R., Eskandar, E. N., Madsen, J. R., Anderson, W. S., Brown, E. N., Halgren, E. & Cash, S. S. (2011), 'Single-neuron dynamics in human focal epilepsy', *Nature neuroscience* **14**(5), 635–641.

Usman, S. M., Khalid, S. & Bashir, S. (2021), 'A deep learning based ensemble learning method for epileptic seizure prediction', *Computers in Biology and Medicine* **136**, 104710.

Usman, S. M., Usman, M., Fong, S. et al. (2017), 'Epileptic seizures prediction using machine learning methods', *Computational and mathematical methods in medicine* **2017**.

Van Rossum, G. & Drake, F. L. (2009), *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA.

Wang, X., Wang, X., Liu, W., Chang, Z., Kärkkäinen, T. & Cong, F. (2021), 'One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial eeg', *Neurocomputing* **459**, 212–222.

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* .

Zabihi, M., Kiranyaz, S., Rad, A. B., Katsaggelos, A. K., Gabbouj, M. & Ince, T. (2015), 'Analysis of high-dimensional phase space via poincaré section for patient-specific seizure detection', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **24**(3), 386–398.

Zarei, A. & Asl, B. M. (2021), 'Automatic seizure detection using orthogonal matching pursuit, discrete wavelet transform, and entropy based features of eeg signals', *Computers in Biology and Medicine* **131**, 104250.