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

William Riddell

April 9, 2024

Word Count: 10,000

Supervised by Kashinath Basu

Faculty of Technology, Design and Environment
School of Engineering, Computing and Mathematics

Contents

1 Abstract **5**

2 Acknowledgements **5**

3 Introduction **5**
- 3.1 Background . . . 5
- 3.2 Aim and Objectives . . . 6
- 3.3 Project Requirements . . . 6

4 Background Review **7**
- 4.1 Datasets . . . 7
- 4.2 Machine Learning (ML) Models . . . 11

5 Project Methodology and Development **16**
- 5.1 Decisions . . . 16
  - 5.1.1 Approach . . . 16
  - 5.1.2 Development System . . . 17
  - 5.1.3 Version Control . . . 17
  - 5.1.4 Software and Libraries Used . . . 18
  - 5.1.5 Schedule . . . 18
- 5.2 Development . . . 19
  - 5.2.1 Electroencephalogram (EEG) Data Extraction . . . 19
  - 5.2.2 Short-Time Fourier Transform (STFT) . . . 19
  - 5.2.3 Notch Filter . . . 20
  - 5.2.4 Synthesize Data . . . 21
  - 5.2.5 Convolutional Neural Network (CNN) . . . 21
- 5.3 Problems Encountered . . . 22
- 5.4 Results . . . 23

6 Professional Issues and Risks **23**
- 6.1 Professional Issues . . . 23
- 6.2 Risks . . . 25
  - 6.2.1 Risk Matrix . . . 25
  - 6.2.2 Risk Assessment . . . 25
- 6.3 Review . . . 27
  - 6.3.1 Convolutional Neural Network (CNN) Architecture . . . 27
  - 6.3.2 Subject Specific Model . . . 27

6.3.3 Subject Generic Model 27
6.3.4 Short-Time Fourier Transform (STFT) Tuning 27
6.3.5 Possible Issues 27
6.3.6 Future Development 27

7 Conclusion **27**

8 Final Thoughts **27**

9 Bibliography **28**

10 Appendix **30**

Acronyms

AI Artificial Intelligence. 5

CHB-MIT Children's Hospital Boston. 16, 18, 19, 30

CNN Convolutional Neural Network. 2, 11, 13–16, 19, 21, 27

DSTL Dynamical Entrainment; difference of short-term Lyapunov exponents. 12, 14, 15

EDF European Data Format. 18, 19

EEG Electroencephalogram. 2, 5, 6, 16, 18–20, 30

EEGs Electroencephalograms. 5, 7, 19

kNN K Nearest Neighbor. 11–13

LR Logistical Regression. 11, 13–16

ML Machine Learning. 2, 5, 6, 9, 11, 13, 16–18, 20, 22, 24

PCA Principal Component Analysis. 22

SPLV Phase-locking Synchrony. 12, 14, 15

STFT Short-Time Fourier Transform. 2, 3, 12, 16–21, 27

SVM Support Vector Machine. 11–15

1 Abstract

asdf

2 Acknowledgements

asdf

3 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)

3.1 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 therefore greatly benefit from a prediction system in their daily life.

3.2 Aim and Objectives

This project will aim to develop an consistent ML model trained to classify either preictal, interictal and ictal periods. The model will have to achieve a high degree of accuracy ($\geq 90\%$) when being applied to EEG data in real-time. Furthermore a real-time simulation will need to be developed along with an ML pipeline and model parameter tuning.

Objectives

1. Research and find an suitable dataset allowing for the classification of the interictal, preictal, and ictal periods.
2. Research and find an suitable ML model approach.
3. Create a data preperation and preprocessing pipeline which extracts the dataset into an ML format suitable for training and testing.
4. Undergo parameter tuning and model architecture tuning of the selected ML model approach..
5. Produce a simulation interface which streams EEG data in real time through the preprocessing pipeline and into the optimal model. The simulation should show the current classification to the user every second.

3.3 Project Requirements

The final product needs to have the ability to accept a stream of raw EEG data, it will need to run the data through a preprocessing pipeline and then through a model. It will need to display the current classification to the user. This process should update at least once a second. The final model is also required to have a classification accuracy of $\geq 90\%$.

4 Background Review

4.1 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

4.2 Machine Learning (ML) Models

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

- Machine Learning (ML) models
  - K Nearest Neighbor (kNN).
  - Support Vector Machine (SVM).
  - Logistical Regression (LR).
  - Convolutional Neural Network (CNN).

- Preprocessing Pipelines
  - Time domain, using the third order Butterworth bandpass filter.
  - Frequency domain using the Fourier transform.
  - Time-Frequency domain using Wavelet Decomposition.
  - Cross-correlation.
  - Non-linear Interdependence.
  - Dynamical Entrainment; difference of short-term Lyapunov exponents (DSTL).
  - Phase-locking Synchrony (SPLV).
  - Entropy of the phase difference.
  - Wavelet Coherence.
  - Short-Time Fourier Transform (STFT)

K Nearest Neighbor (kNN) and Support Vector Machine (SVM)

(Savadkoohi, Oladunni & Thompson 2020) built a feature extraction process that extracted the "time domain" using the Butterworth filter (1-70 Hz) 1, the "frequency domain" using a Fourier transform 2, and the "time-frequency domain" using Wavelet decomposition for the entire dataset. The resulting data is then split into its 5 brain wave bands; Delta, Theta, Alpha, Beta, and Gamma. From these 4 variables were extracted, mean, variance, skewness, and kurtosis, leading to 60 total extracted features.

$$y(n) = \sum_{i=0}^{N} a_i \cdot x(n-i) + \sum_{j=1}^{N} b_j \cdot y(n-j)$$

Figure 1: Third Order Butterworth bandpass filter

$$\widehat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\ e^{-i2\pi\xi x}\,dx.$$

Figure 2: Fourier transform equation

A prediction was calculated from an kNN and SVM model for each separate domain. The results are shown in 6 showing the potential of each model.

Model	Accuracy %		
	TD	FD	T-FD
Support Vector Machine (SVM)	99.5	100	100
K Nearest Neighbor (kNN)	99.5	99	99.5

Table 6: Results from (Savadkoohi et al. 2020) showing results for the Time Domain, Frequency Domain, and Time-Frequency Domain

The classification times however were not included in the report. Extraction in real-time may not be applicable for an approach so extensive in its feature extraction.

Logistical Regression (LR), Support Vector Machine (SVM), and Convolutional Neural Network (CNN)

Mirowski et al. used the (Freiburg 2024) dataset containing intracranial recordings and measured the performance of an LR, SVM, and CNN ML models. They compiled 6 different preprocessing pipelines and tested the selected models against each. The preprocessing pipelines are as follows:

- The Cross-correlation between pairs of EEG channels were calculated with delays ranging from -0.5 to 0.5. The delays allowed for the propagation and processing time of brainwaves. Cross-correlation describes the amount f has to be shifted along the x axis to equal $g3$. Only the maximum value of the cross-correlation values were retained for training.

$$(f \star g)(\tau) \triangleq \int_{-\infty}^{\infty} \overline{f(t)} g(t+\tau)\, dt$$

Figure 3: Cross-Correlation

- Non-linear interdependence “which measures the distance, in state-space, between time-delay embedded trajectories of two EEG channels” was also extracted. This is an bivariate feature which "measures the Euclidian distance, in reconstructed state-space, between trajectories described by two EEG channels".

- The third method was Dynamical Entrainment; difference of short-term Lyapunov exponents (DSTL), where an Lyapunov exponent 4 describes the rate of separation for two trajectories, which in this report was based off "a common measure of the chaotic nature of a signal" (Mirowski, Madhavan, LeCun & Kuzniecky 2009)

$$\lambda = \lim_{t\to\infty} \lim_{|\delta \mathbf{Z}_0|\to 0} \frac{1}{t} \ln \frac{|\delta \mathbf{Z}(t)|}{|\delta \mathbf{Z}_0|}$$

Figure 4: Maximal Lyapunov (λ) exponent

- The last 3 features were Phase-locking Synchrony (SPLV), Entropy, and Coherence of the phase difference. These were extracted from the the wavelet transformation The wavelet transformation extracts the frequency-specific and time-dependent phases.

$$SPLV_{a,b}(f) = \frac{1}{N} \sum_{l=1}^{N} e^{i\left(\phi_{a,f}(l) - \phi_{b,f}(l)\right)}$$

Figure 5: Phase-locking Synchrony (SPLV) extraction from the wavelet transformation

The following results show the number of patients with perfect seizure prediction results "(no false positives, all seizures predicted)" from (Mirowski et al. 2009) for each model, for each preprocessing pipeline.

Cross-Correlation		
LR	CNN	SVM
4	9	4
19%	43%	19%

Table 7: Cross-Correlation results.

Non-Linear interdependence		
LR	CNN	SVM
3	10	5
19%	48%	24%

Table 8: Non-Linear interdependence results.

DSTL
SVM
1
5%

Table 9: Dynamical Entrainment; difference of short-term Lyapunov exponents (DSTL) results.

SPLV		
LR	CNN	SVM
10	13	7
48%	62%	33%

Table 10: Phase-locking Synchrony (SPLV) results.

Entropy of Phase Difference		
LR	CNN	SVM
9	11	7
43%	52%	33%

Table 11: Entropy of Phase Difference results.

Distribution of Wavelet Coherence		
LR	CNN	SVM
11	15	8
52%	71%	38%

Table 12: Distribution of Wavelet Coherence results.

These results show the strength of an CNN when compared to other ML models as it's obtained the highest results across all tables. LR however can also achieve strong results when paired with the correct preprocessing pipeline. Both methods would be applicable for this project assuming that the time from recording to final prediction is less that an second.

Convolutional Neural Network (CNN)

Another CNN approach was implemented by Truong et al. (Truong, Nguyen, Kuhlmann, Bonyadi, Yang, Ippolito & Kavehei 2018) who applied an Short-Time Fourier Transform (STFT) transform 7 on 30 seconds of EEG data. Interference was then removed from the produced matrices by applying an notch filter. The final result was then fed into an CNN producing these results:

No. of seizures	Sensitivity (%)	False Positive Rate (/h)
59	81.4 ± 0.0	0.06 ± 0.00

Table 13: Cross-Correlation results.

5 Project Methodology and Development

5.1 Decisions

5.1.1 Approach

From the datasets discussed in 4.1 the Children's Hospital Boston (CHB-MIT) EEG dataset has been chosen due to the large amount of continuous data, suitable for extracting the preictal period 5. The dataset also has EEG nodes positioned on the subject's scalps, fitting with the use case of this project. Furthermore, the CHB-MIT EEG dataset has produced summary plain-text files for each of the subject's recordings, stating when every seizure began and ended allowing for an straightforward methodology when configuring the EEG tuning and training environment.

Through the analysis of various ML models an CNN will be used for this project. Both (Truong et al. 2018) and (Mirowski et al. 2009) shows the potential CNN have when being used in preictal prediction. The approach described in (Truong et al. 2018) was able to achieve similar or better results in comparison with other preprocessing pipelines and / or ML model configurations while using an less computationally intense preprocessing method. This method was based around an STFT extraction and is described in (Truong et al. 2018). This may allow for the classification to be done in real time, meeting the project's objectives 3.2. To ensure this requirement is met an real-time simulation will be developed, showing the raw EEG data and the final prediction every second.

5.1.2 Development System

Due to the linear nature of this project an Waterfall methodology will be used. Waterfall ensures that previous stages of the software are complete before moving on, this will keep the project on track and moving in an forward direction, crucial for the time-limited aspect of this project. Furthermore, due to the analysis undergone through literature reviews the direction and therefore the goals of the project are clearly defined; this further supports the choice of an Waterfall approach.

5.1.3 Version Control

This project will be under version control through the use of Git, along with GitHub as an remote repository. Both the dataset and the trained models will not be under version control. The use of an version control system and remote repository ensures that if anything goes awry an backup of all commits are stored remotely, securing any progress made in the project. Git also gives the benefit of many features; notably branches and stashing, allowing for developing on ideas, features, and bugs without affecting the "main" code, and merging allows for the combination of said branches.

5.1.4 Software and Libraries Used

Name:	Use case:
Google Cloud	Used to download the CHB-MIT EEG dataset
Keras (TensorFlow)	Used to build, train and test the ML models.
Matplotlib	Used in the real-time simulation to plot the Spectrogram and CHB-MIT EEG data.
MNE	Provides classes for loading EEG data from both CSV and EDF files. Also used for running an Short-Time Fourier Transform (STFT) extraction.
NumPy	Utilized functions and classes to store and manipulate loaded data.
Pandas	Utilized functions and classes to store and manipulate loaded data.
scikit-learn	Provides functions used to produce performance metrics for trained ML models.

Table 14: List of Libraries used

5.1.5 Schedule

The schedule describes the way the Waterfall methodology has affected the project; it shows how it's been broken down into smaller, distinct stages with progress being blocked until the previous stage has been completed. The overlap in March and April is when the models are being tuned, therefore the only active development is on the real-time simulation.

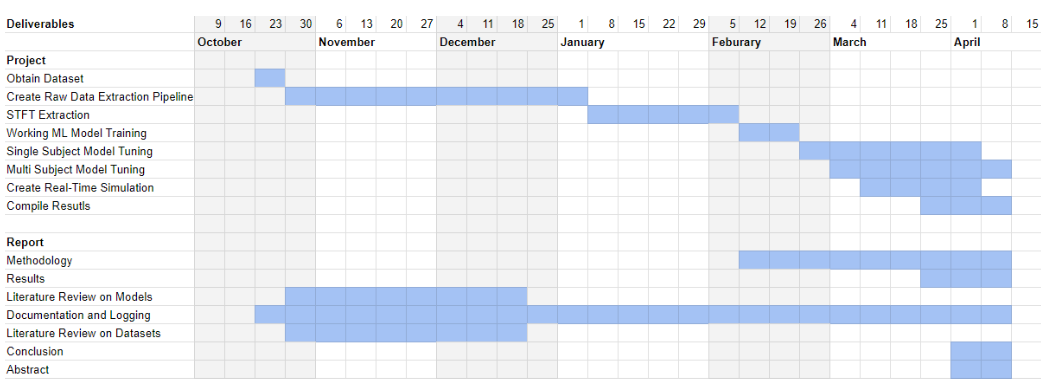

Figure 6: Gantt Chart

5.2 Development

5.2.1 Electroencephalogram (EEG) Data Extraction

The European Data Format (EDF) (Kemp, Värri, Rosa, Nielsen & Gade 1992) is a computer file format that was originally designed for archival and exchange of ... recordings. (Kemp & Roessen 2013). This was the chosen format from the CHB-MIT EEG dataset along with corresponding summary files see 1. From these summary files and the data stored in the EDF files an OOP representation of the dataset was created, containing the channels (individual EEG nodes), recording frequency, and seizure start and end times. With this OOP representation the raw data was extracted from the EDF files. Using the OOP representation the ictal, preictal and interictal periods for each recording were extracted and saved in corresponding directories. Furthermore, only the common channels between all subjects were extracted, following the international 10-20 layout, these were the 17 channel names: FP1-F7, F7-T7,T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, P8-O2, FZ-CZ, CZ-PZ.

5.2.2 Short-Time Fourier Transform (STFT)

For each subject STFT is applied to the data collected from the EEGs to produce an spectrogram of length 30 seconds. These images contains only a single class and were the input into the CNN. STFT can be expressed mathematically;

$$\mathbf{STFT}\{x(t)\}(\tau, \omega) \equiv X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-i\omega t}\, dt$$

Figure 7: Short-Time Fourier Transform (STFT) Window Extraction

Where $w(\tau)$ is an window function and $x(t)$ is the input signal from the EEGs.

This can also be explained as taking an Fourier transform of the EEG signals after an window function has been applied, and then sliding an window across the result. The sliding window transforms the one-dimensional output from the Fourier transform into two-dimensional data allowing for visual analysis.

There are various parameters for an STFT transformation:

Window Function	Used to isolate signal currently undergoing analysis. Optimal functions have low to no artefacts left in the signal and creates no discontinuities at section boundaries.
Window Size	Changes the size of the window function. This affects the resolution of both time and frequency, leading to the uncertainty principal; either variable will be in high resolution. See 8 and 9
Time Step (step size or hop size)	This is the distance between windows. Influences window over or underlap, as well as directly affecting computational load.

Table 15: List of Short-Time Fourier Transform (STFT) parameters

For this project these variables were selected:

Window Function	Sine Window: $w[n] = \sin\left(\frac{\pi n}{N}\right) = \cos\left(\frac{\pi n}{N} - \frac{\pi}{2}\right), \quad 0 \leq n \leq N$.
Window Size	7680. The length of the input data.
Time Step (step size or hop size)	3840. This leads to an window overlap.

Table 16: List of Short-Time Fourier Transform (STFT) parameters choices

See 9 for an Spectrogram produced with these parameters.

5.2.3 Notch Filter

"Power line interference may severely corrupt neural recordings at 50/60 Hz and harmonic frequencies. The interference is usually non-stationary and can vary in frequency, amplitude and phase." (Keshtkaran & Yang 2014) This poses a large issue when training an ML model against EEG recordings as the interference may affect the model's ability to pick up on characteristics, or may mislead the model. 8 clearly shows power line interference with two horizontal bands around 60 Hz and again around 78 Hz. Due to this power line interference needs to be removed, although the time-sensitive nature of this project meant an resource light method was required. Through analysis done by MR Keshtkaran, Z Yang an notch filter has various drawbacks such as not entirely removing the interference (Keshtkaran & Yang 2014), however for this project the alternatives were too computationally intense to run in an real-time setting. Therefore an notch filter was applied to the raw data before STFT was applied.

5.2.4 Synthesize Data

There was also a need to synthesize data in order to fix the class imbalance. For some subjects their ictal periods were very short, only a couple of seconds long in some cases, or the ictal periods began less than 20 minutes into the recording, leading to the preictal period being cut short. Decreasing the number of STFT spectrograms for each class was not a solution as for most subjects their ictal spectrogram count was insufficient, therefore synthesizing spectrogram windows was required to bring the count for each class inline with the interictal class. This was achieved with an sliding window method, see 10. The sliding window offset was calculated for each subject such that each class had the same number of spectrograms as the interictal period.

5.2.5 Convolutional Neural Network (CNN)

For the characteristics stated in ¡¿, an CNN was chosen to classify the spectrogram images. Both training an single model or tuning an selection was implemented. The tuning parameters concerning the architecture was the number of convolutional blocks, number of dense layers, and dense layer size. This allowed an variety CNN with different complexities to be trained and tested. Furthermore the tuning parameters concerned with an individual model was the number of epochs and the batch size. Therefore for each model architecture, an comprehensive view of its ability could be determined. For each model the training and validation accuracy was recorded, along with the final confusion matrix which allowed an F1 score, recall, and precision metric to be generated. The timestamps for model testing and training were also recorded.

5.3 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.

5.4 Results

6 Professional Issues and Risks

6.1 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.

6.2 Risks

6.2.1 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 17: Risk Matrix

6.2.2 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 18: Risk Assessment

6.3 Review

6.3.1 Convolutional Neural Network (CNN) Architecture

asdf

6.3.2 Subject Specific Model

asdf

6.3.3 Subject Generic Model

asdf

6.3.4 Short-Time Fourier Transform (STFT) Tuning

asdf

6.3.5 Possible Issues

asdf

6.3.6 Future Development

asdf

7 Conclusion

asdf

8 Final Thoughts

(Riddell 2023)

9 Bibliography

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*

Freiburg, S. P. P. (2024), 'Eeg database', `https://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database`. Accessed: 2024-04-07.

Kemp, B. & Roessen, M. (2013), 'European data format now supports video', *Sleep* **36**(7), 1111–1111.

Kemp, B., Värri, A., Rosa, A. C., Nielsen, K. D. & Gade, J. (1992), 'A simple format for exchange of digitized polygraphic recordings', *Electroencephalography and clinical neurophysiology* **82**(5), 391–393.

Keshtkaran, M. R. & Yang, Z. (2014), 'A fast, robust algorithm for power line interference cancellation in neural recording', *Journal of neural engineering* **11**(2), 026017.

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.

Riddell, W. (2023), *Test Book*, British London.

Savadkoohi, M., Oladunni, T. & Thompson, L. (2020), 'A machine learning approach to epileptic seizure prediction using electroencephalogram (eeg) signal', *Biocybernetics and Biomedical Engineering* **40**(3), 1328–1341.

Truong, N. D., Nguyen, A. D., Kuhlmann, L., Bonyadi, M. R., Yang, J., Ippolito, S. & Kavehei, O. (2018), 'Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram', *Neural Networks* **105**, 104–111.

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* .

10 Appendix

Listing 1: Example Summary file from the CHB-MIT EEG dataset.

```
File Name: chb06_03.edf
File Start Time: 03:09:42
File End Time: 7:09:42
Number of Seizures in File: 0

File Name: chb06_04.edf
File Start Time: 07:09:51
File End Time: 10:50:52
Number of Seizures in File: 2
Seizure 1 Start Time: 327 seconds
Seizure 1 End Time: 347 seconds
Seizure 2 Start Time: 6211 seconds
Seizure 2 End Time: 6231 seconds

File Name: chb06_05.edf
File Start Time: 10:51:20
File End Time: 14:51:20
Number of Seizures in File: 0

File Name: chb06_06.edf
File Start Time: 14:51:23
File End Time: 18:51:23
Number of Seizures in File: 0

File Name: chb06_07.edf
File Start Time: 18:51:31
File End Time: 22:51:31
Number of Seizures in File: 0

File Name: chb06_08.edf
File Start Time: 22:51:39
File End Time: 26:51:39
Number of Seizures in File: 0

File Name: chb06_09.edf
File Start Time: 02:51:47
```

```
File End Time: 6:51:47
Number of Seizures in File: 1
Seizure 1 Start Time: 12500 seconds
Seizure 1 End Time: 12516 seconds

File Name: chb06_10.edf
File Start Time: 06:51:54
File End Time: 10:51:54
Number of Seizures in File: 1
Seizure 1 Start Time: 10833 seconds
Seizure 1 End Time: 10845 seconds

File Name: chb06_12.edf
File Start Time: 14:52:10
File End Time: 18:52:10
Number of Seizures in File: 0

File Name: chb06_13.edf
File Start Time: 18:52:20
File End Time: 22:52:20
Number of Seizures in File: 1
Seizure 1 Start Time: 506 seconds
Seizure 1 End Time: 519 seconds

File Name: chb06_14.edf
File Start Time: 22:52:35
File End Time: 26:52:35
Number of Seizures in File: 0

File Name: chb06_15.edf
File Start Time: 02:52:43
File End Time: 6:52:43
Number of Seizures in File: 0

File Name: chb06_16.edf
File Start Time: 06:52:51
File End Time: 7:43:21
Number of Seizures in File: 0

File Name: chb06_17.edf
```

```
File Start Time: 07:45:51
File End Time: 11:45:51
Number of Seizures in File: 0

File Name: chb06_18.edf
File Start Time: 11:45:55
File End Time: 13:58:03
Number of Seizures in File: 1
Seizure 1 Start Time: 7799 seconds
Seizure 1 End Time: 7811 seconds

File Name: chb06_24.edf
File Start Time: 08:23:24
File End Time: 12:23:24
Number of Seizures in File: 1
Seizure 1 Start Time: 9387 seconds
Seizure 1 End Time: 9403 seconds
```

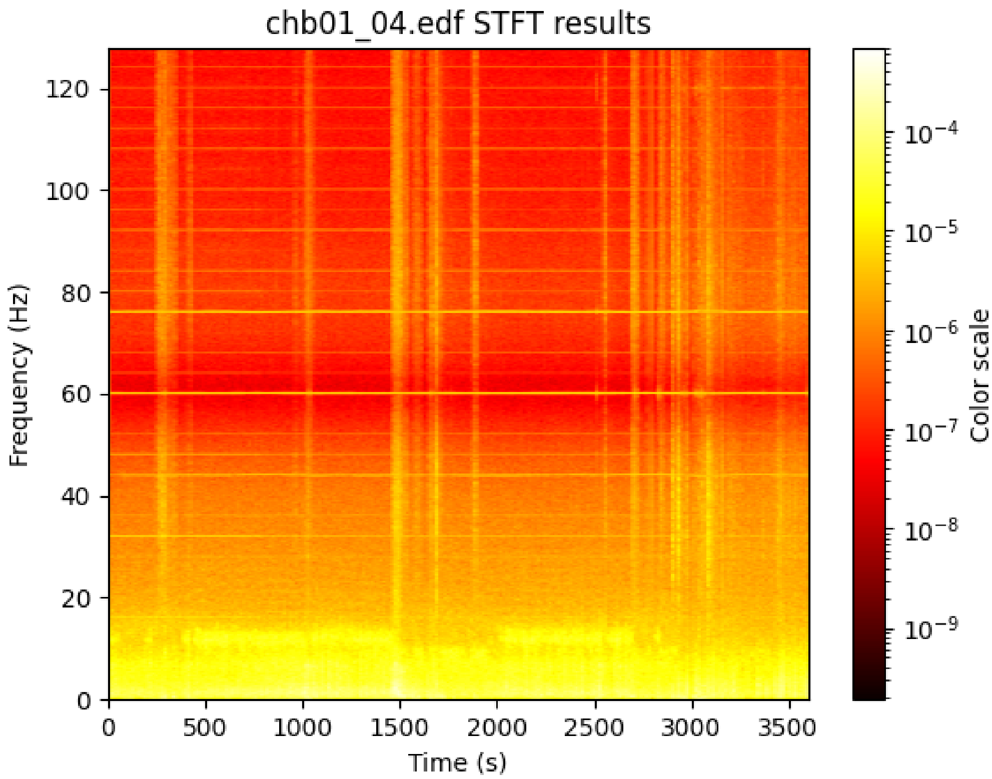

Figure 8: An STFT window with high frequency resolution.

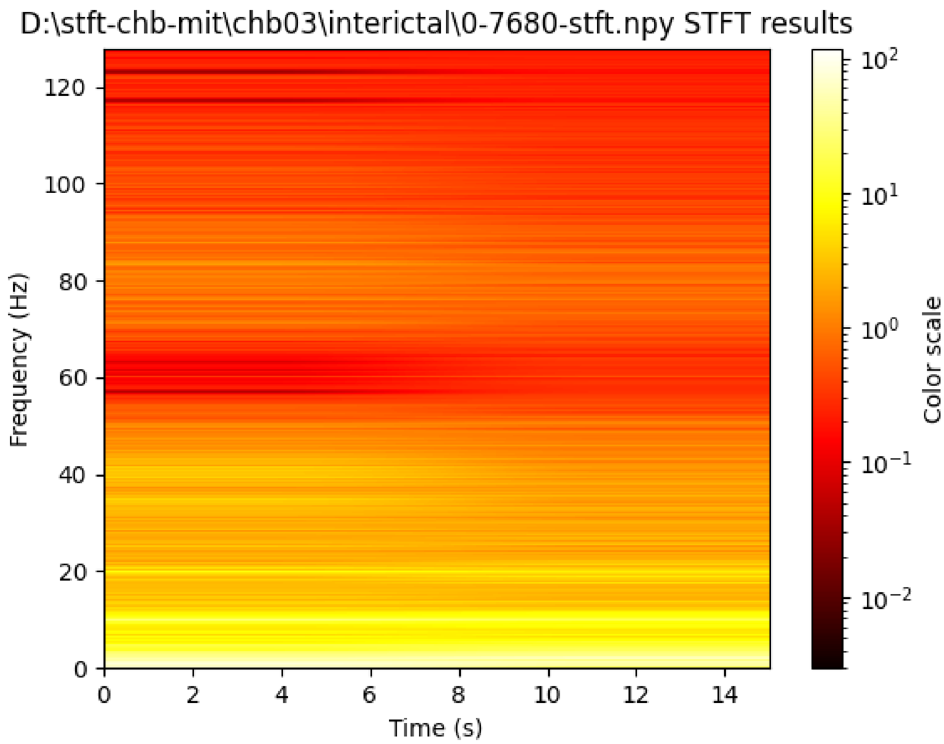

Figure 9: An STFT window with high time resolution.

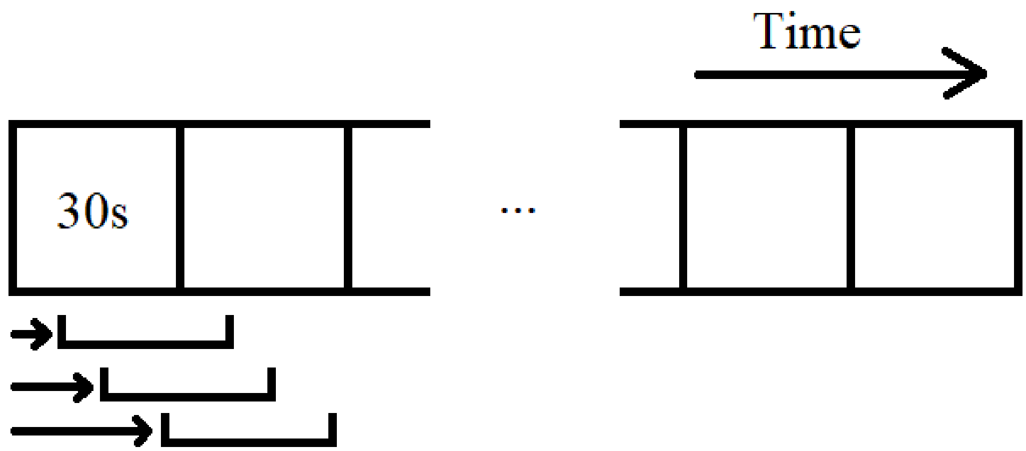

Figure 10: Sliding Window technique for synthesizing data spectrogram images