Executive Summary

For this report, I will be investigating the methodologies of which to best model Epileptic Seizure occurrences from electroencephalogram (EEG) readings with regards to the computational resources available to me. With the goal of achieving high measurable success with model when compared to baseline.

For medical devices, an accuracy ratio of 3:1 is typically an accepted ratio by most certification bodies. [1] So, the passing criteria for model accuracy is set at 75%. Accuracy will be taken into consideration; however, other metrics take prevalence when the occurrence of a seizures is not as common as not having one. According to the dataset, the seizures are recorded 25% of the time, as seen in Figure 1, where class 0 represents the number of cases without epileptic seizures and class 1 representing the number of cases with epileptic seizures.

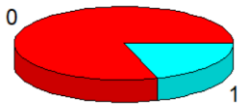


Figure 1 - class allocation

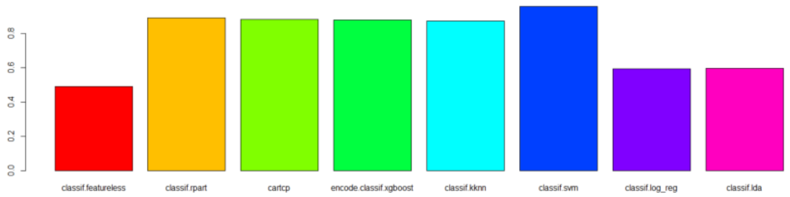


Figure 2- classic model accuracy

Misclassification of epileptic seizures will also be taken into consideration. There are two different cases of misclassification when dealing with this dataset, wrongly identifying an EEG as a seizure case or wrongly identifying an EEG as a non-seizure case. One of misidentifications is more important. In a financial and moral context, the misclassification of a seizure case as a non-seizure case is worse; this is due to complications that can arise from long lasting epileptic convulsions which when not treated can cause permanent brain damage or even death, which will in turn have legal and financial ramifications through law suits or fines . [2] According to MedicalNegligenceAssist.co.uk, severe brain damage can award a maximum of £275,000. [3] While for the case of misclassifying a patient as having a seizure when he/she is not, a visual check by the first responder to the scene can mitigate further resources being used, so there are minimal concerns for this case. The metric recall takes this into consideration, therefore, the metrics to obtain the model best suited for seizure classification are recall and accuracy, with recall taking priority over accuracy.

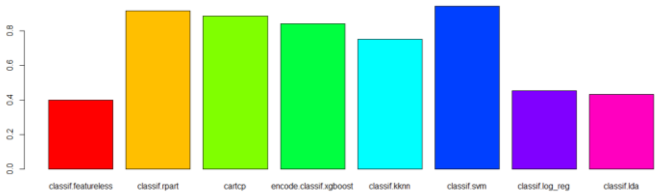


Figure 3 – classic machine learning recall

The models in figure 2 and 3 can be developed quickly and have achieved great results. The best performing model, SVM (blue bar), has attained the highest accuracy of 95.5% and the high recall rate at 94.1% . To boost the results, I investigated stacking the top 2 performing models, SVM and rpart.

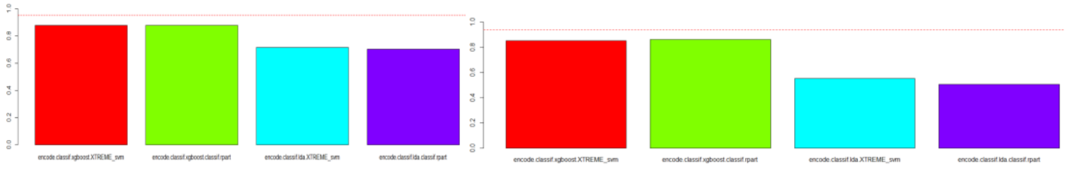


Figure 4- Stacked models accuracy(left) , Stacked model recall rate(right)

The red line in figure 2 and 3 represent the scores from the best performing model, SVM, from Figure 1 at 95.5%. As can be seen in Figure 2 and Figure 3, none of the stacked models are able to outperform the plain SVM model.

Next, several configurations of multi-layer perceptron (MLP) were investigated and the best model achieved better performance than the SVM.

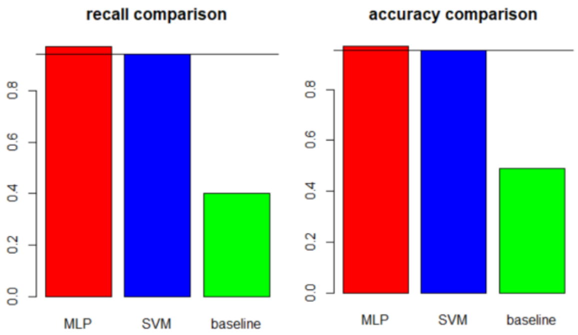


Figure 5- MLP vs SVM , recall comparison(left), accuracy comparison(right)

As visible from Figure 5, the MLP out-performs the SVM on in both metrics of accuracy and recall rate. It should be noted that transfer learning techniques were also investigated, time invested in feature engineering and training the deep learning model made this technique unviable as the room for improvement was not worth the cost of additional effort. When comparing the MLP to the baseline model, there is a great margin of improvement, there is a gain of 48% in accuracy and 57.1% gain in recall rate. I believe the goal has been met and therefore suggest the MLP model would be the best model for this application.

Technical Summary

Data Exploration

Each sample is made up of 178 data points, the data points represent the value of the EEG recordings for a second in time. The original 5 Classes were ‘eyes open’, ‘eyes closed’, ‘tumor identification’ , ‘EEG from tumor location’ and ‘seizure activity’. As per dataset description, all the classes except ‘seizure activity’ did not have epileptic seizure activity, this allows us to investigate the occurrences of seizures as a binary classification problem.

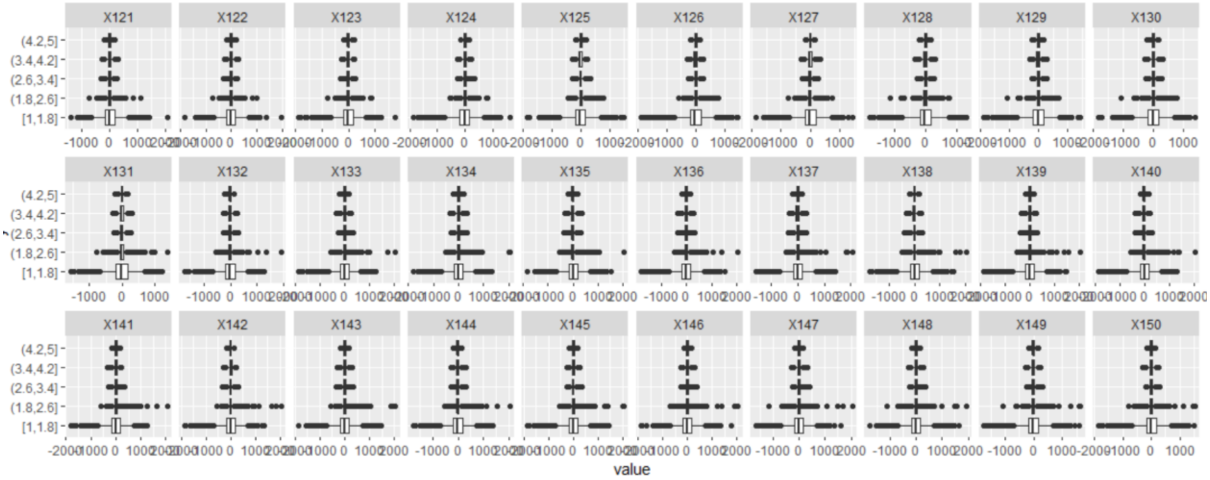


Figure 6- Box and whiskers plot of original dataset

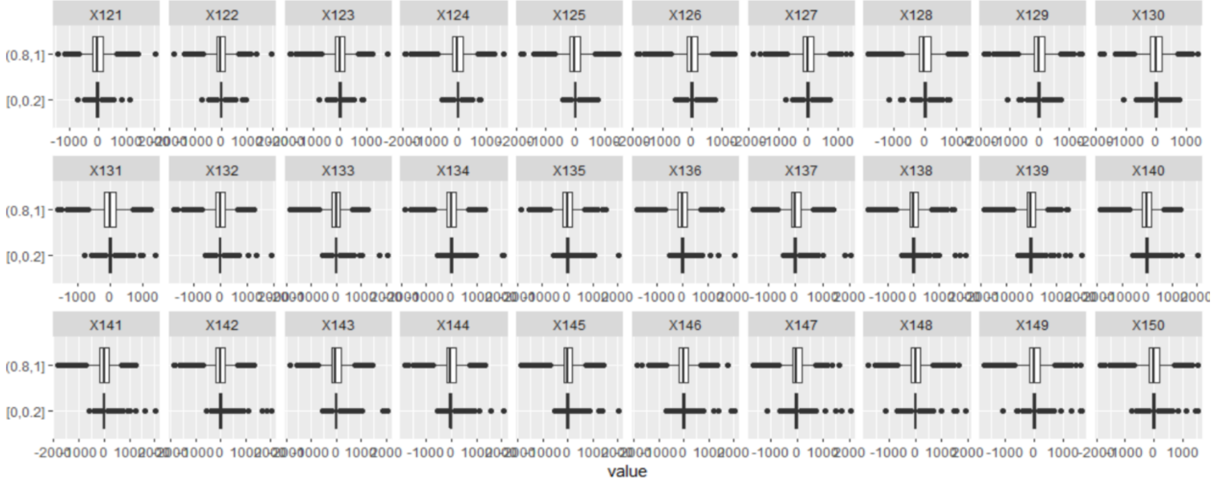


Figure 7 - Box and whisker plot of fused non-epileptic seizure classes

To convert the dataset into a binary classification problem, I converted classes 2-5 into class 0. As can be observed in Figure 6, classes 2-5 have similar distributions while class 1 has a spread out distribution of values for its features. I believe this makes it features of each class easily distinguishable and would not require high model complexity.

Machine learning

Model	Classification error (%)	Classification accuracy (%)	False positive rate (%)	False negative rate (%)	Recall (%)
Baseline model	51.0	49.0	40.0	60.0	40.0
Recurring part	11.0	89.0	13.3	8.6	91.4
Cart	12.0	88.0	12.4	11.5	88.5
XGBoost	12.3	87.7	8.6	16.0	84.0
Knn	12.7	87.3	0.6	24.8	75.2
Svm	4.4	95.5	3.1	5.9	94.1
Logistic regression	40.6	59.4	26.6	54.5	45.5
LDA	40.3	59.7	23.7	56.8	43.2

For the benchmarking of the models, the models used a cross-validation fold of 5 as to have the 80:20 ratio for training and testing for each fold.

Stacking

By stacking models on top, one another, it is possible to get a boost in performance. Used LDA as LDA is a dimensionality reduction algorithm which produce eigenvalues that can help model performance. Similarly, XGBoost does feature selection which reduces the dimensionality to be fed into a feeder model.

Model	Classification Error (%)	Classification Accuracy (%)	False Positive Rate (%)	False Negative Rate (%)
XGBoost + svm	12.0	88.0	11.5	12.5
XGBoost + cart	12.1	87.9	15.5	11.8
LDA + svm	29.3	70.7	11.9	46.6
LDA + cart	30.5	69.5	8.5	52.7

Multilinear Perceptron

Uneven Classes

With it comes to classification problems, a variation of sample counts per class can a negatively influence the fitting of the model. Subsampling the training data can mitigate the issue.

	Class 0	Class 1
Original allocation of classes	6864	1761
After down sampling	1761	1761

I used down sampling to avoid repeats to many repeats of class 1, which might make the model overfit to the reoccurring samples.

Early Stopping

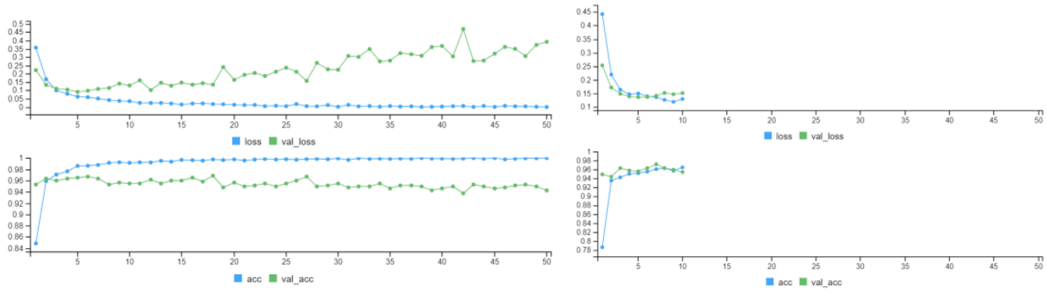


Figure 8- fitting plots , without early stopping (left), with early stopping (right)

Initial investigation shows that training the model on 50 epochs too much, and results in overfitting, as seen in the loss plot the validation loss is continuously going up after the initial 5 epochs. Using Early stopping will mitigate this, for a better generalizing model.

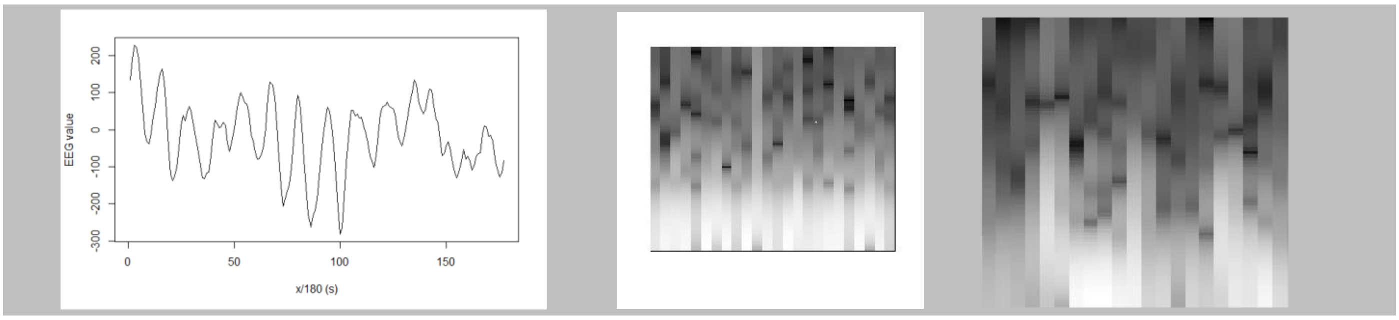
Trial no.	No. of dense Layers	activation	No of units	Dropout layers	Epochs	accuracy	precision	recall	F1 score
1	2	Relu, x 2	89,45	0	50	96.4	98.8	96.6	97.7
2(w early)	2	Relu x 2	89,45	0	9	96.7	98.8	97.0	97.9
3	2	Leaky relu x 2	89,45	0	9	96.8	98.9	97.1	98.0
4	2	Leaky relu x 2	89,45	1(rate=0.5)	10	96.7	99.0	96.9	97.9
5	2	Parametric relu x 2	89, 45	1(rate=0.5)	9	96.0	98.6	96.4	97.5
6	2	Parametric Relu x 2	89, 45	2(rate=0.5)	11	97.0	99.4	96.9	98.1
7	2	Leaky relu x 2	89,45	2(rate=0.5)	15	97.0	99.0	97.1	98.1
8	2	Relu x2	89,45	2(rate=0.5)	12	97.0	99.1	97.1	98.1
9	3	Relu x 3	89,45,22	3(rate=0.5)	19	96.3	99.1	96.3	97.7
10	2	Relu x 2	89,45	1(rate=0.5)	10	96.5	98.8	96.8	97.8

Using an inverse pyramid scheme, with the input layer as the largest size, I chose the next hidden layer to be half the size and if there is another layer after that, it would be half the size of the previous layer. I have also used dropout rates at 50% as that is the usual dropout rate.

Transfer Learning

Transfer learning is the practice of utilizing pre-trained models for a different task from which it was originally trained for. It allows users to leverage on the pre-trained weights and deep models to save on time and experimentation.

I decide to investigate Image Classification using transfer learning. To attain the image dataset required to train the model, I converted the signals into spectrograms.



On the left is the original signal. In the middle is the spectrogram plot with the axis and labels removed. On the right is the cropped image with the white background removed.

For the image dataset creation, I initially took more than 5 hours creating the dataset. As I was constantly prototyping the pipeline, this put a major barrier in my timeline. I have found that creating the images in batches of 50 has resulted in faster overall creation. Rather than converting the dataset in its entirety. Overall, transfer learning seems like over-kill as simpler models achieved great success, unless the dataset grows much bigger, deep learning would not be necessary.

Model	epochs	Layers unfrozen	Optimizer	modifications	Accuracy
inceptionV3	10	0	Rmsprop		50.4
	10	138	Sgd(lr = 0.0001)		0.456

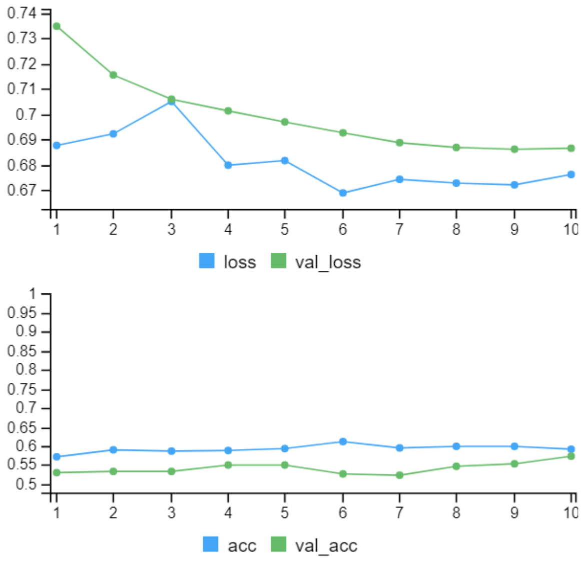


Figure 9 - Unfreezing 2 inception cells, the equivalent of 138 layers.

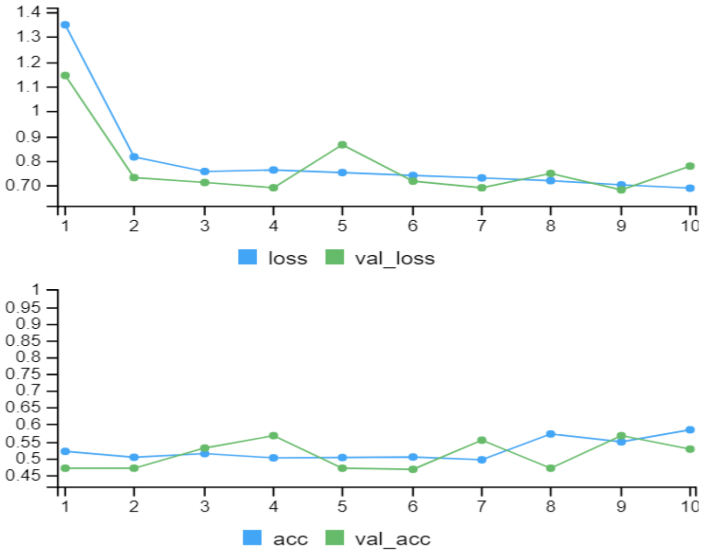


Figure 10 - only additional layers were unfrozen

References

[1] A. Joshi, “Calibration requirements in ISO 13485,” 8 Mar 2019. [Online]. Available: https://advisera.com/13485academy/blog/2019/03/08/calibration-requirements-in-iso-13485/.

[2] “Status epilepticus and emergency treatment,” Epilepsy action, May 2019. [Online]. Available: https://www.epilepsy.org.uk/info/firstaid/emergency-treatment-seizures-last-long-time.

[3] “How Much Could My Medical Negligence Claim Be Worth?,” MedicalNegligenceAssist, [Online]. Available: https://www.medicalnegligenceassist.co.uk/how-to-claim/how-much-could-my-medical-negligence-claim-be-worth/. [Accessed Mar 2019].