Seizure Forecast of Epilepsy Patients Using Electroencephalogram (EEG)

EEG data is quite often used for diagnosis and treatment of patients with epilepsy. They are referred to the brainwave activities.

Four Stages

Preictal
prior to seizure

Post-ictal
after seizures

Interictal
between seizures
or baseline

Ictal
seizure

Task

classification for interictal and preictal EEG.

Dataset

We have interictal and preictal EEG data from 15 electrodes of one patients

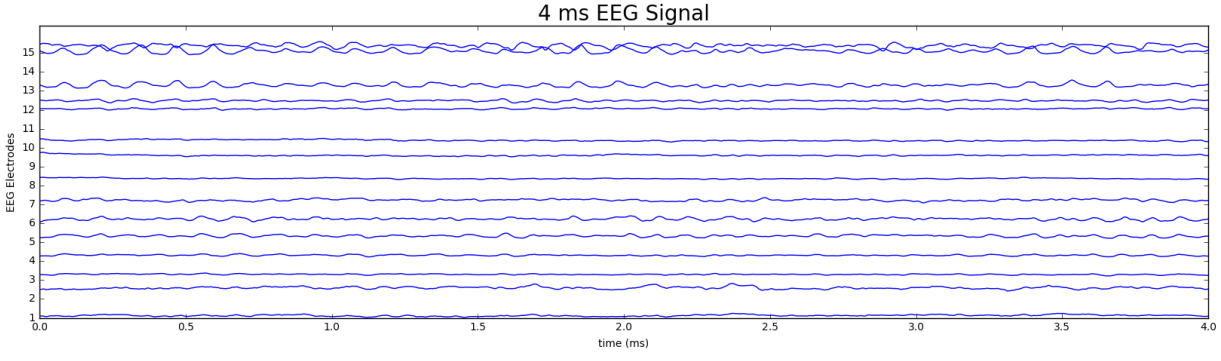
Duration: 500mins of interictal, 180 minutes of preictal
Sampling frequency: 5kHz
Segment: one segment to be 4ms (20000 samples), in total 7500 interictal segments, 2700 preictal segments

There are EEG data from 15 electrodes. For each segment of EEG data, we want to extract some features from both time and frequency domain.

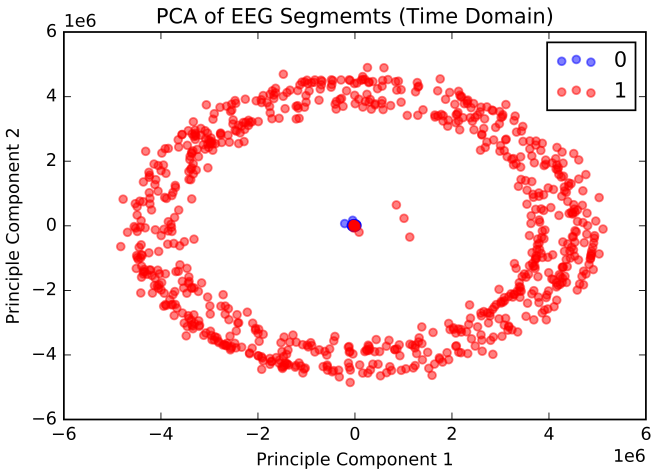
The first plot is a typical interictal EEG signal of 4 ms.

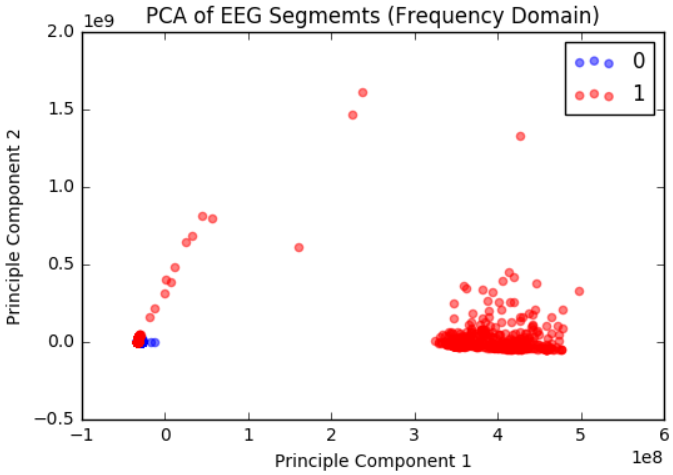
From the second plot, we can see difference of both time and frequency domain.

4ms EEG signal of 15 electrodes



Attempt 1: PCA to reduce dimension



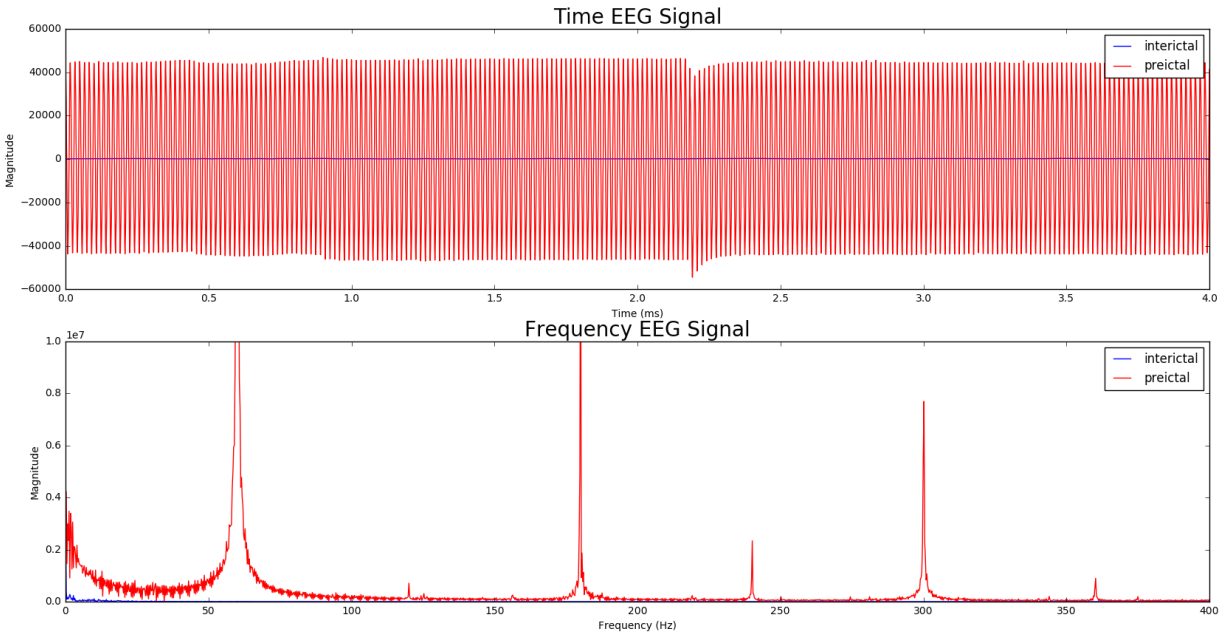


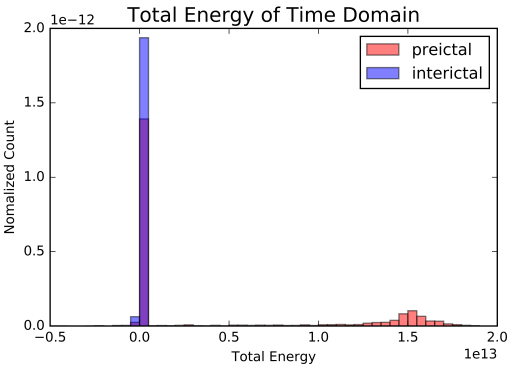
Class	Accuracy
Preictal	0.3
Interictal	1

Class	Accuracy
Preictal	0.3
Interictal	1

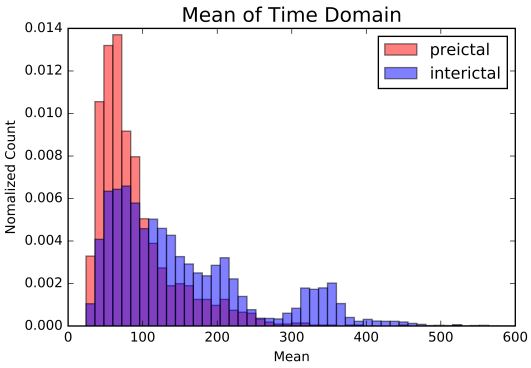
We try to do PCA to reduce dimension of the segmented signals (4ms each). Although we see good separation of both time and frequency domain, there are still too many overlapped data of both classes. Then we decide to get rid of PCA and run classification models on extracted features.

Attempt 2: Extract Features both from time and frequency domain

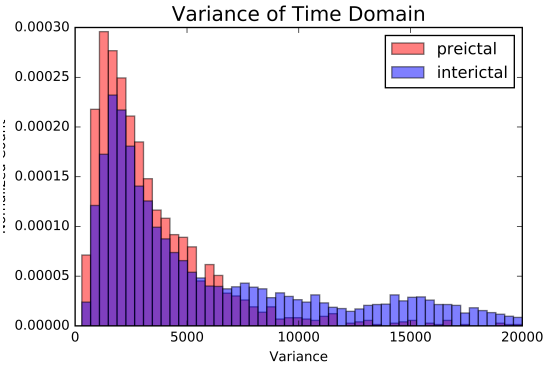




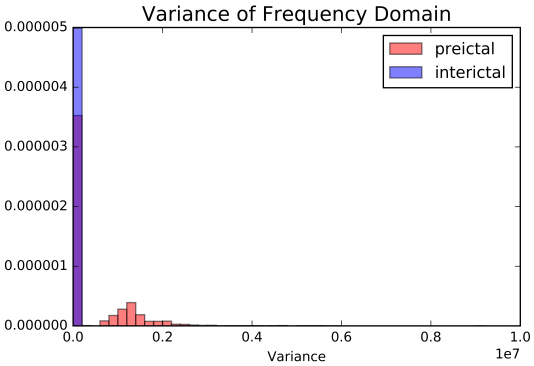
- Sum of squared magnitude in time domain



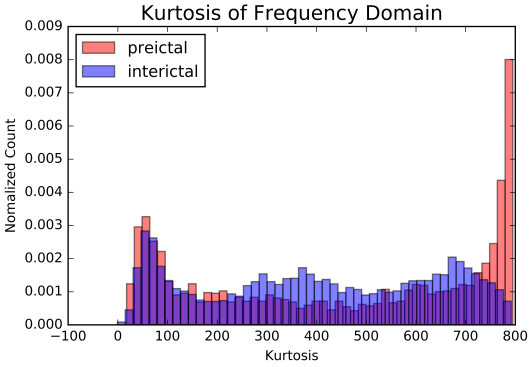
- Mean of absolute magnitude of time domain



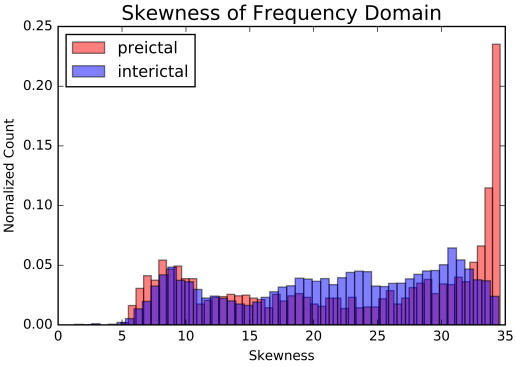
- Variance of absolute



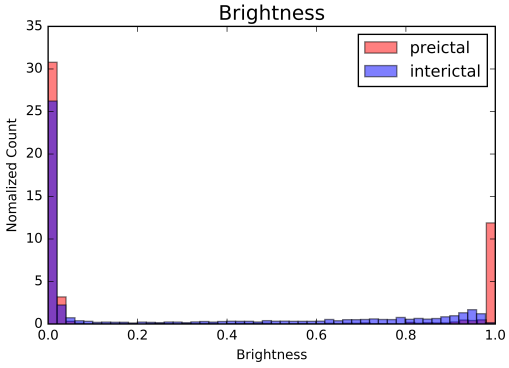
- Variance of spectrum magnitude, upper limit = 200Hz domain



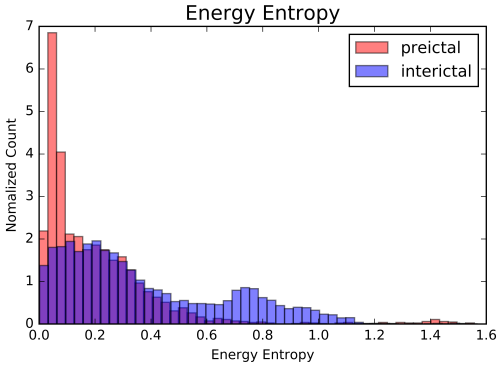
- Fourth standardized



- Third standardized moment domain

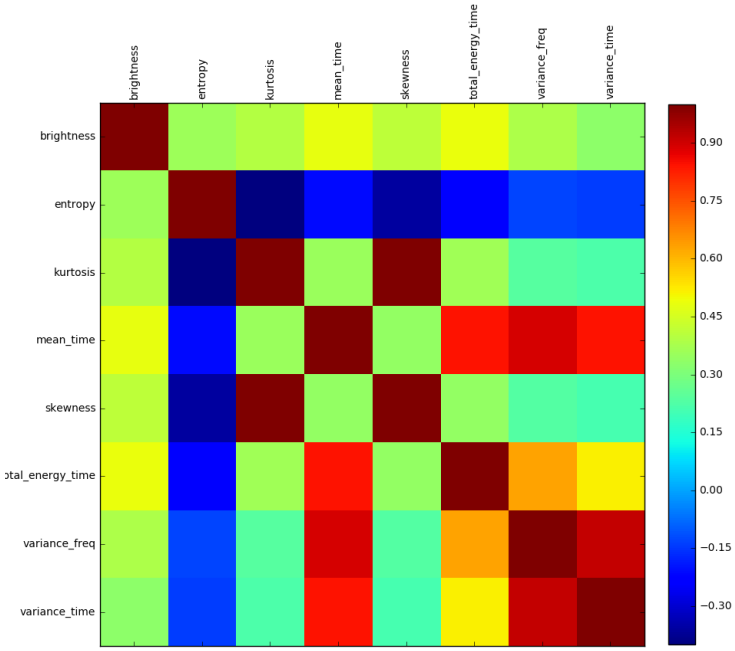


- Percentage of energy above



- 0Hz, 0-10Hz, 10-50Hz, 50-100Hz, 100-200Hz, 200-300Hz, 300-400Hz, 400-2500Hz.

Correlation Matrix



This is the correlation matrix of 8 features for one electrode.

Most of the features are not heavily correlated. But we see some obvious correlation between the last three features: total energy in time, variance of spectrum and variance in time and time mean magnitude.

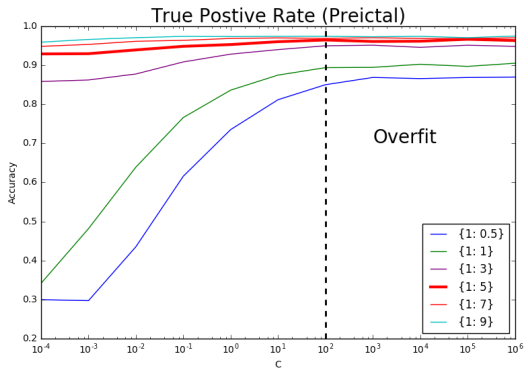
We will keep all the features in the

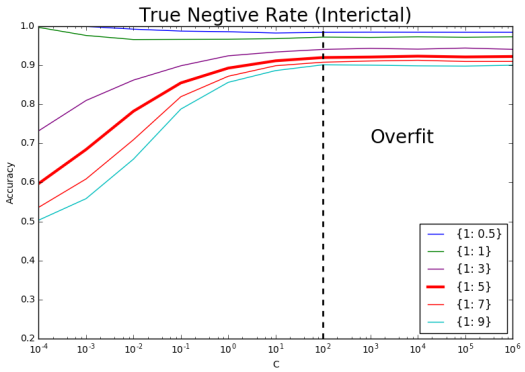
Try Classification Models on Extracted Features

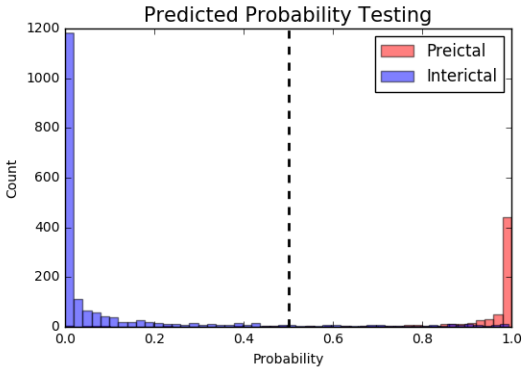
Dataset: 120 features (8 * 15), 2700 interictal and 7500 preictal data
Train and Test: split the data to have 75% training set and 25% testing set
Standardization: All the features are numeric, standardize all the predictors
Cross validation: Do 5 folds on the training set to tune and select parameters.
Accuracy: Report ROC curve, accuracy of both classes (true positive rate and true negative rate) and total accuracy on the testing set.

Logistic Regression

Best parameters: class weight = 1:5, C = 100







Tune two parameters: class weight and regularization parameter C.

From the plot above, we visualize the relationship between tuning and accuracy of each class. We plot accuracy against different C, with each line representing different class weight.

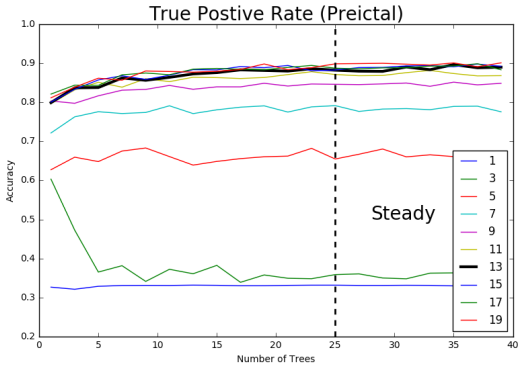
As C increases, the model tends to overfit the training set. And as weights increase, the influence becomes less obvious.

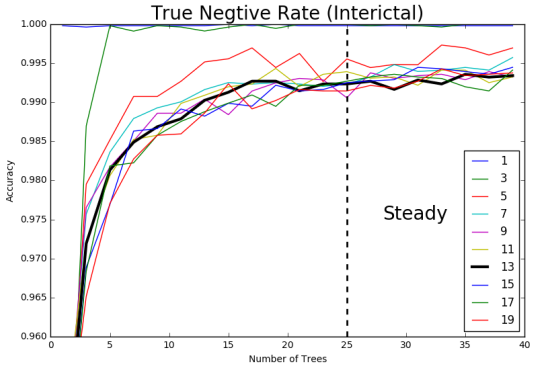
In the end we select the thick red line to be our best parameters. While we achieve good accuracy on both classes.

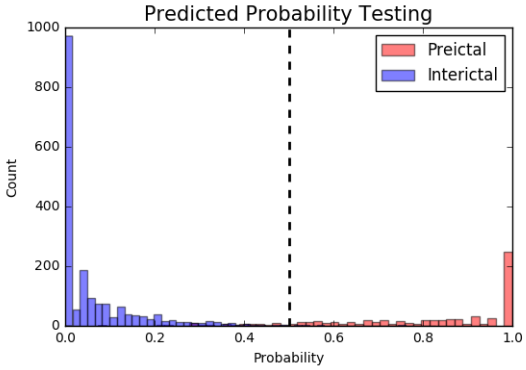
Also, from left plot, we see good separation on predicted probability of both interictal and preictal data.

Random Forest

Best parameters: n_estimator = 25, max_depth = 13







Tune two parameters: n_estimator (number of trees) and maximum depth.

From the plot above, we visualize the relationship between tuning and accuracy of each class. We plot accuracy against different number of trees, with each line representing different maximum depth.

As n_estimator increases, the accuracy becomes steady (can not reduce variance more by combining more trees) . And as maximum depth increase, the influence becomes less obvious.

In the end we select the thick black line to be our best parameters. While we achieve good accuracy on both classes.

Also, from left plot, we see good separation on predicted probability of both interictal and preictal data.

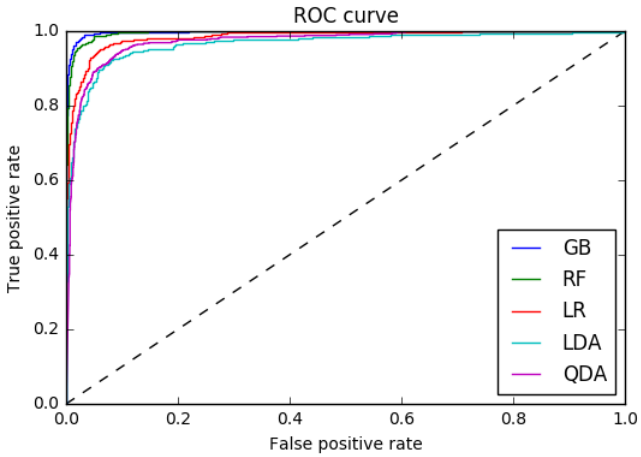
Other models

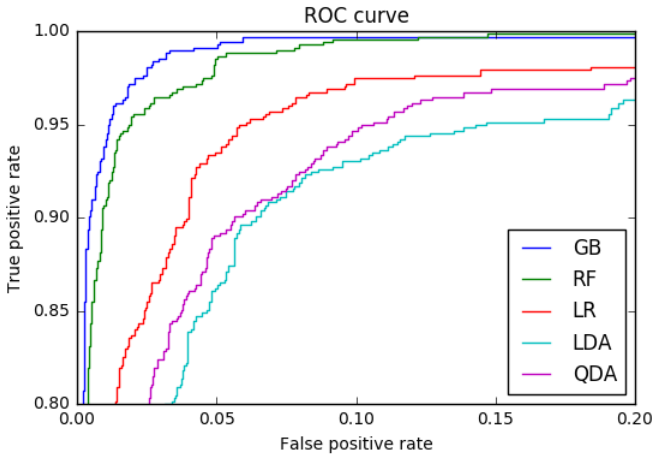
LDA: Best parameters: priors = [0.3,0.7]
QDA: Best parameters: priors = [0.3,0.7]
KNN: Best parameters: n_neighbors = 1
Gradient boost: Best parameters: max_depth=4,n_estimators=300

Best predictors

The **top three predictors** we find from logistic regression and random forest are: frequency variance, time mean magnitude and total time energy. These predictors are highly correlated from the correlation matrix.

ROC analysis on Testing Set





We successively run logistic regression, random forest, LDA, QDA, KNN, and boosting and tuned the best parameters on the training set. Now we compare the ROC curve of these models. All the models are doing really good job.

The best model is boosting. The ROC curve hang to the upper left corner, achieving both highest true positive rate and lowest false positive rate.

Accuracy Comparison on Testing Set

Accuracy	KNN	Logistic Regression	LDA	QDA	Random Forest	Boosting
Preictal	0.88	0.82	0.79	0.96	0.97	0.97
Interictal	0.94	0.98	0.97	0.81	0.96	0.98
Total	0.93	0.94	0.92	0.85	0.97	0.98

So far, using the features we extracted from both interictal and preictal EEG signal of the same patient, we have achieved, at highest, total classification accuracy of 0.98, with 0.97 accuracy on preictal data and 0.98 accuracy on interictal data.

Conclusion and further work

If we can have access to EEG data from many more different features, we can try to give a general seizure forecast metric. Also, we can explore on extracting more features from the EEG signal.