Definition

Project Overview

This project will use the confusion EEG dataset hosted on Kaggle at https://www.kaggle.com/wanghaohan/eeg-brain-wave-for-confusion. This dataset consists of EEG data taken from students watching MOOC courses of varying difficulty. EEG (or electroencephalogram) is a method of monitoring brain activity by using external electrodes placed on the subject's scalp that can measure voltage fluctuations from large numbers of neurons firing. While these are not precise enough to measure individual neurons, by 'listening' to the brain as a whole they can get a measurement of the frequencies present. Modern neural models hypothesize that the frequency that neurons fire encode messages, so we may be able to determine information about what the subject is thinking about by looking at those frequencies. In particular the device used by the researchers is measuring frequencies in the range of Delta waves (waves with a frequency of 1-3 Hz, or oscillations per second), Theta waves (4-7 Hz), Alpha waves (8-11 Hz), Beta waves (12-29 Hz), and Gamma waves (30-100 Hz). The device also reports a few proprietary measurements supposedly related to paying attention and meditation, along with the raw signal. Finally we have labels for each of the videos. The predefined label is whether the researchers expected the subject to be confused or not, and the self defined label is the self reported level of confusion. Both of these are reported as binary 0s or 1s.

Problem Statement

The purpose is to identify signals from the EEG that indicate whether or not the student is confused by the subject matter. In theory a confusing subject matter should require additional concentration, or at least a different type of focus from the student, which may be observable in the EEG data. This could be used in a product providing some sort of computer adaptive educational content with a "consumer" level EEG device (which are becoming more and more available). For instance if it detects that the student is confused, it can slow down the material, provide some more background material, or just give the student more time to consider what they are learning. In contrast, if the student shows low levels of confusion, it can be more confident in moving forward. Computer adaptive education is not a new concept, but typically involves a lot of time consuming testing to see if the student has learned the material. While this probably would not eliminate the need of testing, it could certainly reduce it. It could also give the product more confidence in its result as it would help identify times when the user may know just enough to answer the questions in the exam, but still has some underlying confusion. And it could help evaluate the quality of the learning materials by determining whether or not they are clear to the students.

To achieve this we will train a model using a gradient boosting classifier against the collected EEG to correctly classify whether or not the student reported that they were confused. If after training it can classify the sessions with an acceptable accuracy, then this model could be incorporated in a product to help determine the confusion level of the user. Furthermore by looking at what features it is using we can focus further research on the brain wave frequencies that our model identifies.

Metrics

To test any models I build with this data I will compute the accuracy score. The data is very balanced, with 51% of the sessions involving a confused student and 49% involving a non-confused student. At this early stage, since I do not have an exact product planned out, I do not have a bias towards preferring either false negatives or false positives, so accuracy should be an effective metric. I will also look at the Brier score as it will also look at the probability the model computes by computing what is effectively the mean squared error between the predicted probability of the positive label and the label itself. A product using these models could take these probabilities to judge how confident it is on the prediction, thus it makes sense to evaluate how well they perform as well.

Analysis

Data Exploration

If we look at the first five rows in table 1 we can get a feel of what the data looks like. There is one entry per time period per video per student, each with the values of each of the collected statistics at that time period along with the predefined and self defined labels.

Table 1

	subject ID	Video ID	Attention	Meditation	Raw	Delta	Theta	Alpha 1	Alpha 2
0	0	0	56	43	278	301963	90612	33735	23991

1	**0**	**0**	**40**	**35**	**-50**	**73787**	**28083**	**1439**	**2240**
2	**0**	**0**	**47**	**48**	**101**	**758353**	**383745**	**201999**	**62107**
3	**0**	**0**	**47**	**57**	**-5**	**2012240**	**129350**	**61236**	**17084**
4	**0**	**0**	**44**	**53**	**-8**	**1005145**	**354328**	**37102**	**88881**

	Beta 1	Beta 2	Gamma1	Gamma2	predefined label	Self-defined label
0	**27946**	**45097**	**33228**	**8293**	**0**	**0**
1	**2746**	**3687**	**5293**	**2740**	**0**	**0**
2	**36293**	**130536**	**57243**	**25354**	**0**	**0**
3	**11488**	**62462**	**49960**	**33932**	**0**	**0**
4	**45307**	**99603**	**44790**	**29749**	**0**	**0**

We can also look at some summary statistics of the data in table 2.

Table 2

	count	mean	std	min	25%	50%	75%	max
subject ID	12811.0	4.487394	2.865373	0.0	2.0	4.0	7.0	9.0
Video ID	12811.0	4.390602	2.913232	0.0	2.0	4.0	7.0	9.0
Attention	12811.0	41.313871	23.152953	0.0	27.0	43.0	57.0	100.0

Meditation	12811.0	47.182656	22.655976	0.0	37.0	51.0	63.0	100.0
Raw	12811.0	65.570760	597.921035	-2048.0	-14.0	35.0	90.0	2047.0
Delta	12811.0	605785.261728	637623.562614	448.0	98064.0	395487.0	916623.0	3964663.0
Theta	12811.0	168052.602919	244134.569620	17.0	26917.5	81331.0	205276.0	3007802.0
Alpha 1	12811.0	41384.350636	72430.815187	2.0	6838.0	17500.0	44779.5	1369955.0
Alpha 2	12811.0	33183.393178	58314.100751	2.0	6852.0	14959.0	34550.5	1016913.0
Beta 1	12811.0	24318.368980	38379.684967	3.0	6140.0	12818.0	27406.0	1067778.0
Beta 2	12811.0	38144.330263	79066.056294	2.0	7358.5	15810.0	35494.0	1645369.0
Gamma1	12811.0	29592.552806	79826.366922	1.0	4058.0	9763.0	24888.0	1972506.0
Gamma2	12811.0	14415.972992	36035.232415	2.0	2167.5	5116.0	12669.5	1348117.0
predefined label	12811.0	0.470377	0.499141	0.0	0.0	0.0	1.0	1.0
Self-defined label	12811.0	0.512606	0.499861	0.0	0.0	1.0	1.0	1.0

For these purposes I will group the data by session (one subject and one video) and use the values of the brainwaves in each of the reported ranges. I will not be using the raw data as it doesn't look like there is enough resolution in the data to see individual brain waves on their own. I will also not use the proprietary measurements. They are likely computed as a function of the rest of the data, and therefore not likely useful on their own. Furthermore without knowing how they are computed we would not be able to use them with different equipment. Once aggregated, they look like Figure 3:

Table 3

		Delta	**Theta**	**Alpha 1**	**Alpha 2**	**Beta 1**	**Beta 2**	**Gamma1**	**Gamma2**	**Self-defined label**

subject ID	Video ID									
0	**0**	2723077	1031826	556251	234589	186966	171258	222111	84108	0
	1	3224853	826317	304340	221773	139036	131248	141042	84001	1
	2	3958185	961497	400302	123180	221211	145414	164217	66255	1
	3	2581211	1698512	251577	236024	174228	176850	227196	112579	0
	4	2757383	1011493	167859	90579	95771	144309	181573	49188	0
	5	3059285	1330245	850147	396815	231739	192808	340048	138218	1
	6	2927619	1811549	143813	122927	89965	100927	129326	75859	1
	7	1942380	858788	242933	61527	67659	128280	131320	41331	0
	8	2505972	1812829	327389	219596	98671	139751	184525	106159	1
	9	3529287	1040266	188482	174153	146567	122511	133563	48166	0

For the label, I will use the self defined label, since I feel that is more trustworthy. The predefined labels are making assumptions about how much the students know or how easily they are confused.

Exploratory Visualization

The values of each wavelength over time are shown in figures 1-8 for subject zero during videos 0 (during which he was not confused) and 1 (during which he was confused).

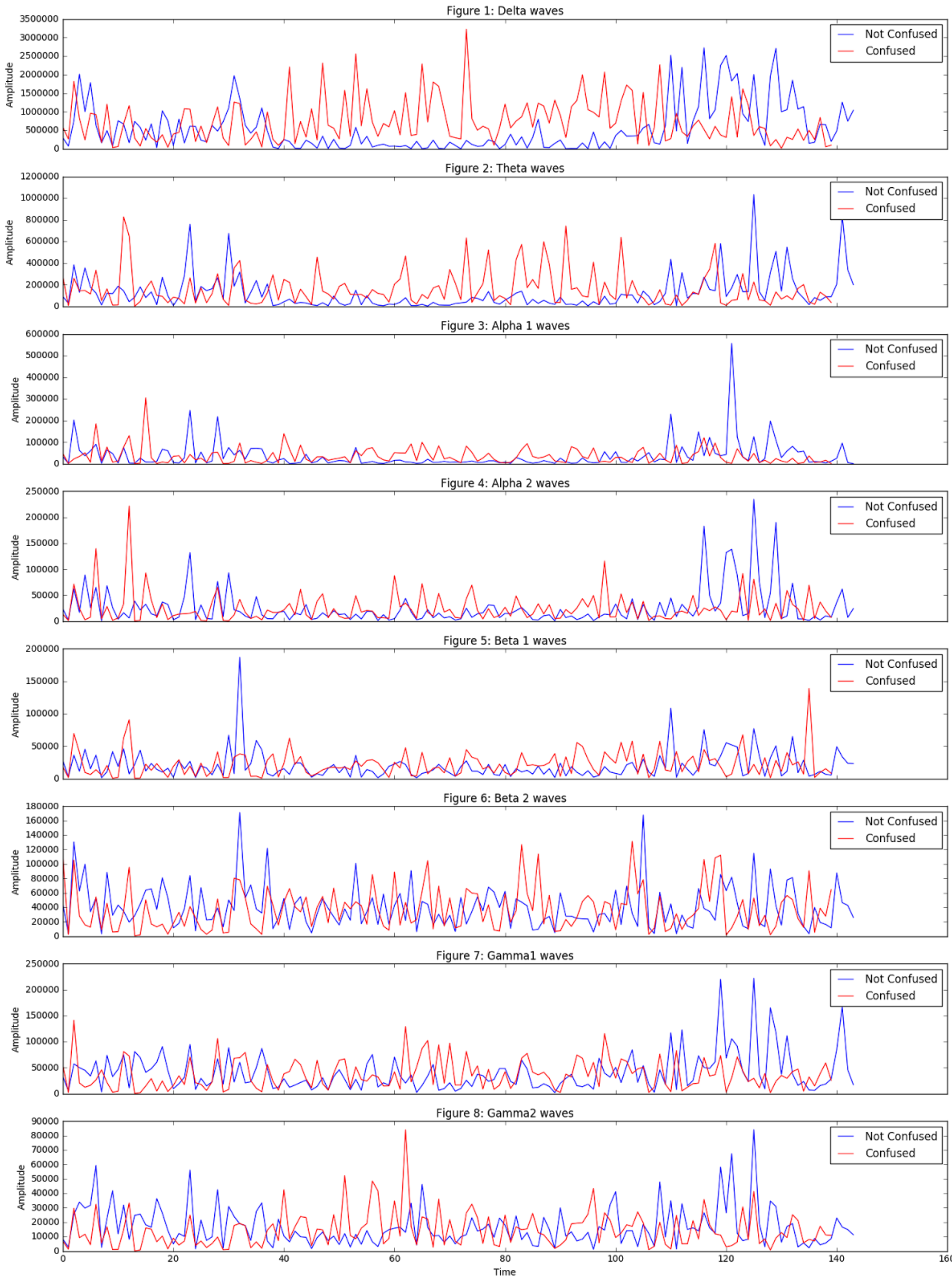


Figure 1: Delta waves
Figure 2: Theta waves
Figure 3: Alpha 1 waves
Figure 4: Alpha 2 waves
Figure 5: Beta 1 waves
Figure 6: Beta 2 waves
Figure 7: Gamma1 waves
Figure 8: Gamma2 waves
Not Confused
Confused
Amplitude
Time

Here we can see some potential differences. Several of the wavelengths do seem to be increased while the subject was watching the confusing video, particularly in the lower wavelengths. The difference seems to be greater in the middle of the video, which could mean either there were times being recorded before and after the video, or there was a "warm up" and "cool down" period for each video during which the subject's brain was no longer considering the video. However, its also possible the higher activity at the end of the lesson that the students understood is a valid signal.

Algorithms and Techniques

To model the brain waves we will use a gradient boosting ensemble classifier. These models tend to be robust against overfitting, which given our limited data is a danger. They can also model multiple paths to a result, which may be necessary as "confusion" is a vague concept which may show itself in multiple ways. It is also possible to interpret them to see what features it is using, which can give us both insight in how the brain is working and can validate the model against existing theories of neural activity.

In terms of features, we have a couple of different ways to extract features from this data. We will aggregate the data for each student/video session to get 100 different data points. From each data point, we can take both the mean and standard deviation of the frequencies. The mean will give a measurement of how much activity in that frequency is occurring, while the standard deviation will give us a measurement if it is staying in a certain level or varying over the video. We can also split the data into different time boxes. From looking at the visualizations above, there is little difference in the early part of the video, which makes sense as the video would not have had enough time to have an impact on the subject's brain. There also looks to be distinct phases in the middle of the video and at the end. Again, this intuitively makes sense as in the middle of the video the student's brain would be concentrating on understanding what is being presented, while at the end they are more reflecting on what has been said.

Benchmark

The researchers providing the dataset report that an accuracy of 65% ends up being "quite decent", so I will try to improve on that. However I do not expect to be able to do too much better, as I do not expect this dataset to be sufficiently complete in its description of the subject's brain to get perfect (or even near perfect) accuracy. Further complicating the matter is that the reporting is rather subjective. The predefined confusion level is not a particularly accurate measurement because it is based on the researcher’s preconceptions of what the subjects know. The self reported confusion level is better, but still is limited in that it depends on the user's internal metric for confusion. There are also perhaps some honesty issues involved, as a student might believe he or she should not be confused on a certain subject even if he or she is, and thus falsely report a 0. With those considerations, I will consider a good model to be one getting above 70% accuracy. We will also look at their brier score to determine how much confidence we can get from the probabilities they are reporting. Finally we will look at how well the models generalize across different students and across different videos.

Methodology

Data preprocessing

The data needs to be aggregated by each student/video. We will pull out their means and standard deviations. We will then separate each session into thirds and find both the means and standard deviations for the middle and last third of the videos. While each wavelength exists on a different scale with average values of 14k for gamma 2 waves to 606k for delta waves, the algorithm we are using is implemented with decision trees, so normalization or removal of outliers is not needed.

Implementation

First we need a function to evaluate how well a given model performs against a given set of data. We will run cross validation scoring the accuracy and the brier score of the model. We also need to be able to test how well each the model generalizes across different students and different videos. To do this we will use LabelKFold folds where the label is either the student id or the video id.

With that given, if we run a gradient tree boosting classifier we find the cross validated model gets an average accuracy of 0.747917 and an average Brier score of 0.197387. For reference a Brier score will always be between 0 and 1, so 0.2 is pretty good. If we evaluate the model across different students, we get a mean accuracy of 0.58 with a standard deviation of 0.172047. Across different videos we get a mean accuracy of 0.72 and a standard deviation of 0.107703. The values for each student are shown in Figure 9, the values for each video are shown in Figure 10.

Figure 9: Accuracy Across Students With Means

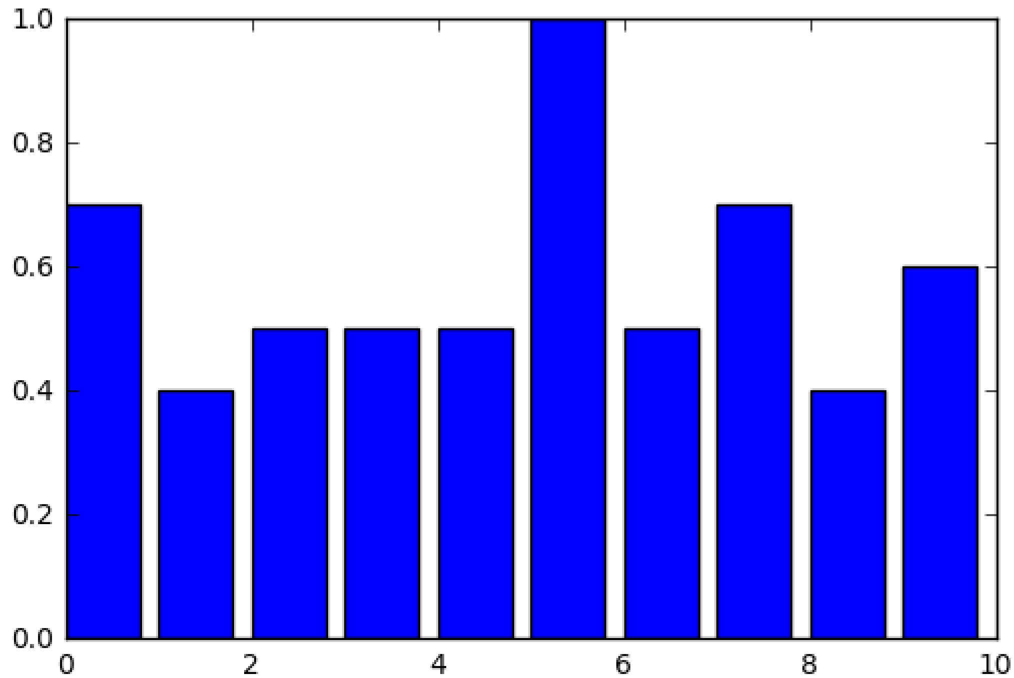
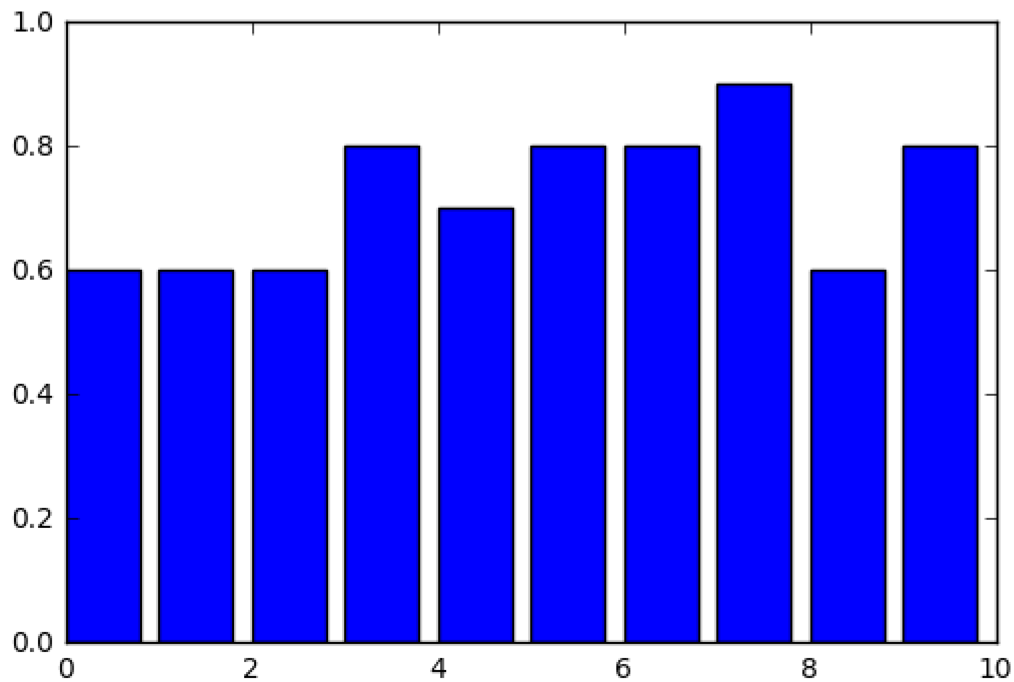


Figure 10: Accuarcy Across Videos With Means



It generalizes pretty well across videos, with the worst accuracy of 60% and several at 80 or 90%. It does less well across students, where for 6 students it did 50% or worse.

If we look at how important the model is treating each feature in table 4, it appears to be focusing primarily on the low frequency delta waves, with theta waves coming in second. Previous research has suggested that Theta waves are involved in confusion, so this is promising. It also corresponds to the visualization we did earlier where the delta and theta waves were noticeably higher in the confused subjects.

Table 4

	Importance
Average Delta	0.209421
Average Theta	0.174372
Average Alpha 2	0.148682
Average Gamma2	0.124358
Average Gamma1	0.108760
Average Alpha 1	0.083931
Average Beta 1	0.077750
Average Beta 2	0.072724

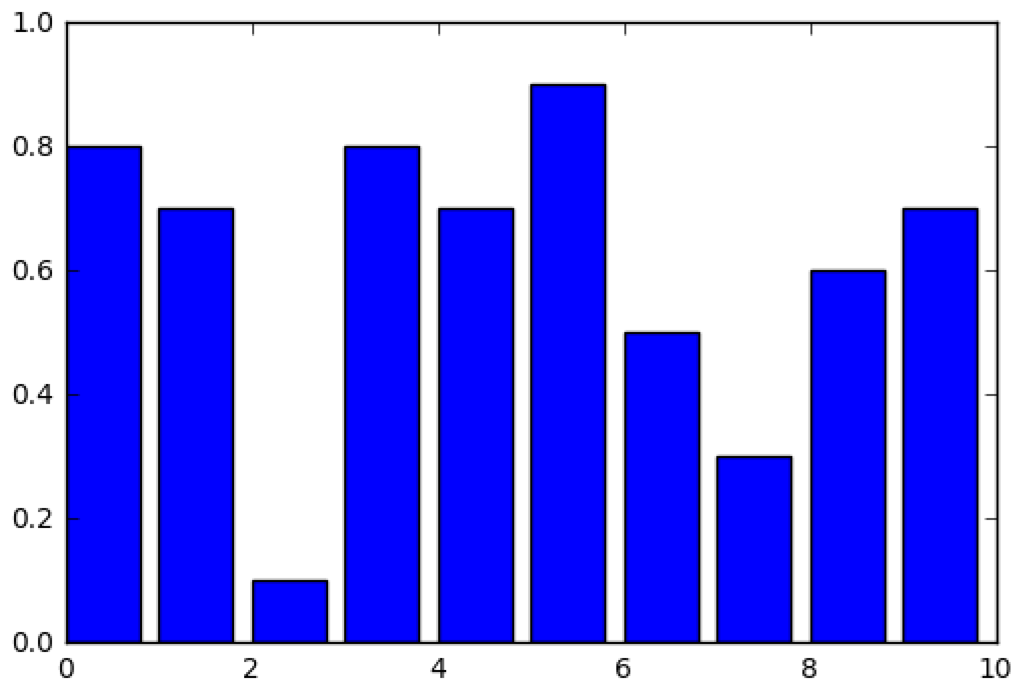
Refinement

The first area to refine the models is in the features they are using. We started out with the mean values of each brainwave group, but we can also include the standard deviations in case the variability of one or more of the features is important. We can also look at the middle and late periods in the video separately.

First we can create a model by adding the standard deviations to the features. With them added, we get an accuracy of 0.725000 and a Brier score of 0.212757. So the standard deviation data does not seem to help. In fact the model performs slightly worse in both accuracy and the Brier score when they are added.

If we use the data separated out into the both mid video and late video as features, we get an accuracy of 0.760417 and a Brier score of 0.185906. So the model appears to be getting slightly better. On average these models seem to be generalizing better both against videos (Figure 12) with a mean accuracy of 74% and students with a mean accuracy of 61%, however across students (Figure 11) it varies quote a bit. While for 6 students it did 70% or better, for one student it is correct only 30% of the time, and for another it is correct only 10 percent of the time.

Figure 11: Accuracy Across Students With Time Separated Means



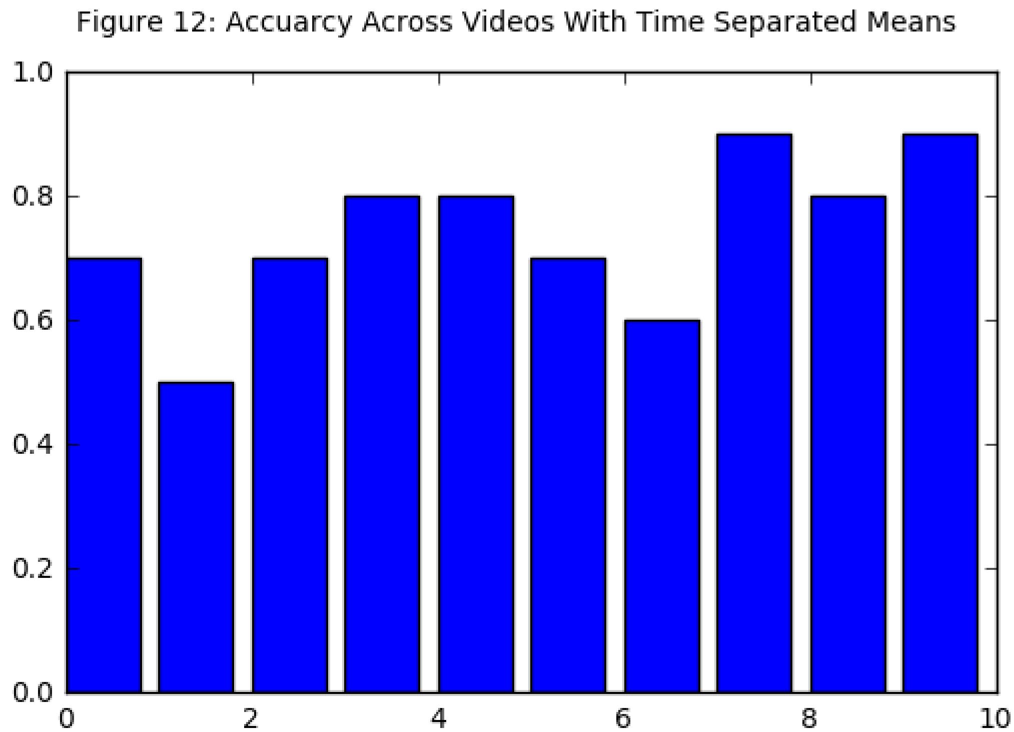


Table 5 shows the importance of each feature in this model. The most important features are the theta and upper alpha waves during the middle of the video. The top scoring waves during the latter third of the video are delta waves, but they are quite a bit lower than the theta and alpha 2 waves.

Table 5

	Importance
Mid Theta Mean	0.183342
Mid Alpha 2 Mean	0.178386
Late Delta Mean	0.082603
Mid Gamma2 Mean	0.078203
Late Beta 1 Mean	0.071403

Mid Beta 1 Mean	0.068555
Mid Alpha 1 Mean	0.066921
Mid Delta Mean	0.051486
Late Gamma2 Mean	0.051451
Late Theta Mean	0.031990
Late Beta 2 Mean	0.030997
Late Alpha 2 Mean	0.025839
Late Alpha 1 Mean	0.025178
Late Gamma1 Mean	0.020886
Mid Beta 2 Mean	0.017127
Mid Gamma1 Mean	0.015634

Since it's possible we are getting some false signals at the end of the video (for instance if the video has completed), let's also look at just the middle part. With that change it is doing noticeably better in terms of both accuracy (0.783333) and the Brier score (0.179587). In terms of it generalizing, across videos (Figure 14) it is generalizing quite well, with 4 videos showing 90% accuracy, and all but one show 70% or above. In terms of students (Figure 13) it is doing better than any of the other feature sets, though still worse than across the videos.

Figure 13: Accuracy Across Students With Middle Means

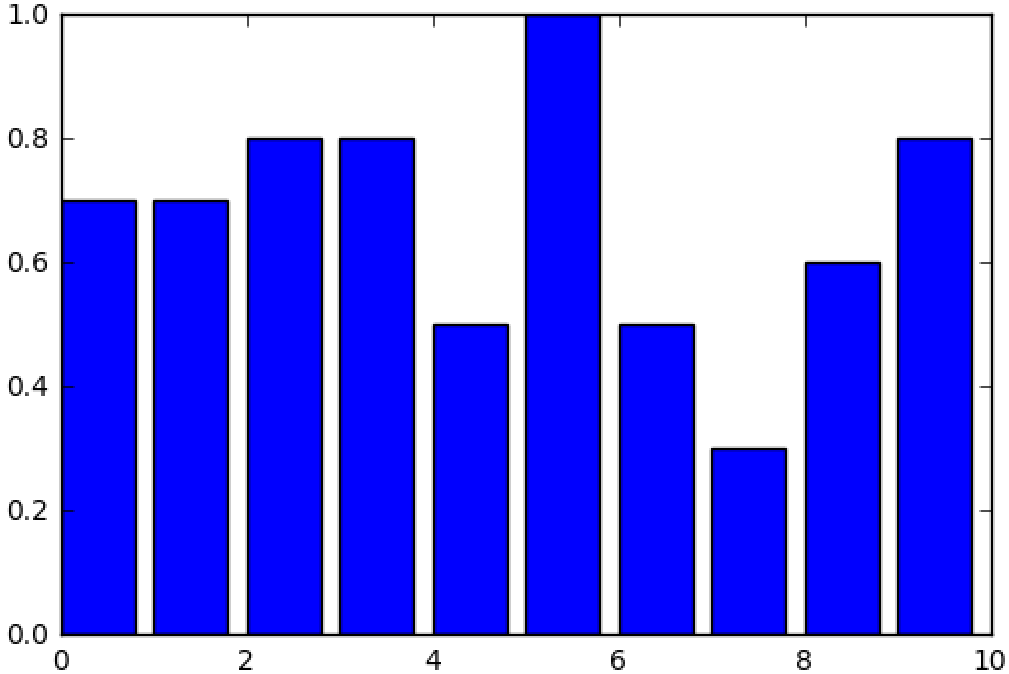
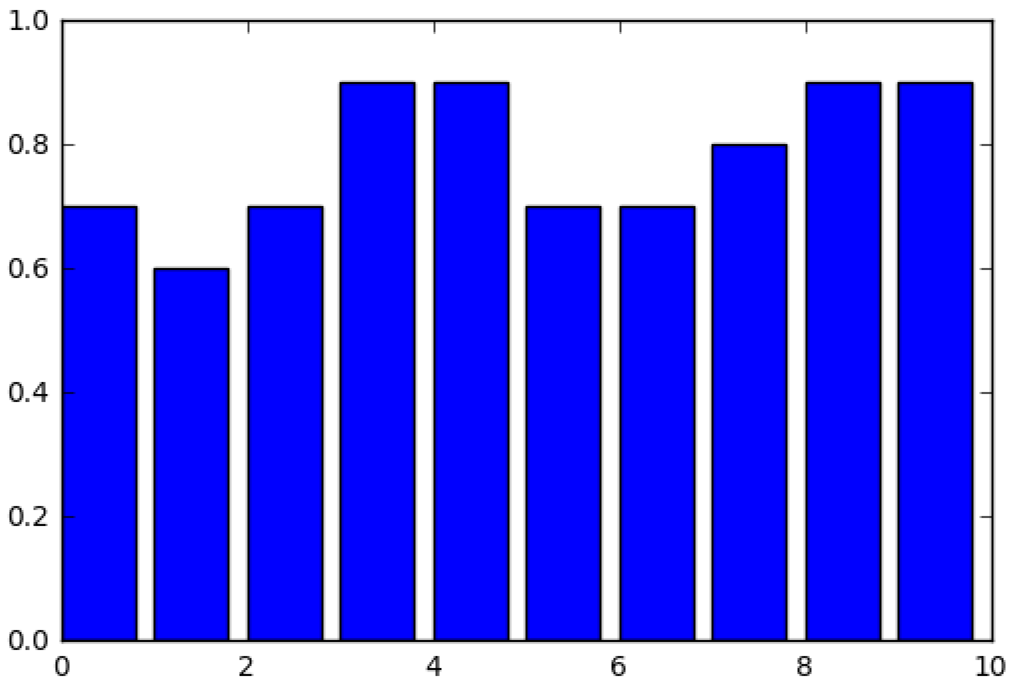


Figure 14: Accuarcy Across Videos With Middle Means



If we look at what features it's using (Table 6), the upper alpha and theta waves show to be the most important. Again, theta waves are expected as previous research has suggested it should be important. The importance of alpha waves is somewhat surprising, especially since there is a clear difference between upper and lower alpha waves

Table 6

	Importance
Mid Alpha 2 Mean	0.215272
Mid Theta Mean	0.203108
Mid Gamma2 Mean	0.152900
Mid Delta Mean	0.103543
Mid Beta 1 Mean	0.097387
Mid Gamma1 Mean	0.082917
Mid Beta 2 Mean	0.081064
Mid Alpha 1 Mean	0.063809

We can now try looking at the middle video means combined with their standard deviations. This doesn't change the accuracy (0.787500) or Brier score (0.170951) much overall, but it is now a bit more consistent across students (Figure 15) with a mean accuracy of 70%. There are now only two students which are getting below 70% accuracy. Overall video accuracy (Figure 16) is slightly lower (at 77%), though now only 4 of them have an accuracy below 0.8

Figure 15: Accuracy Across Students With Middle Means and STDs

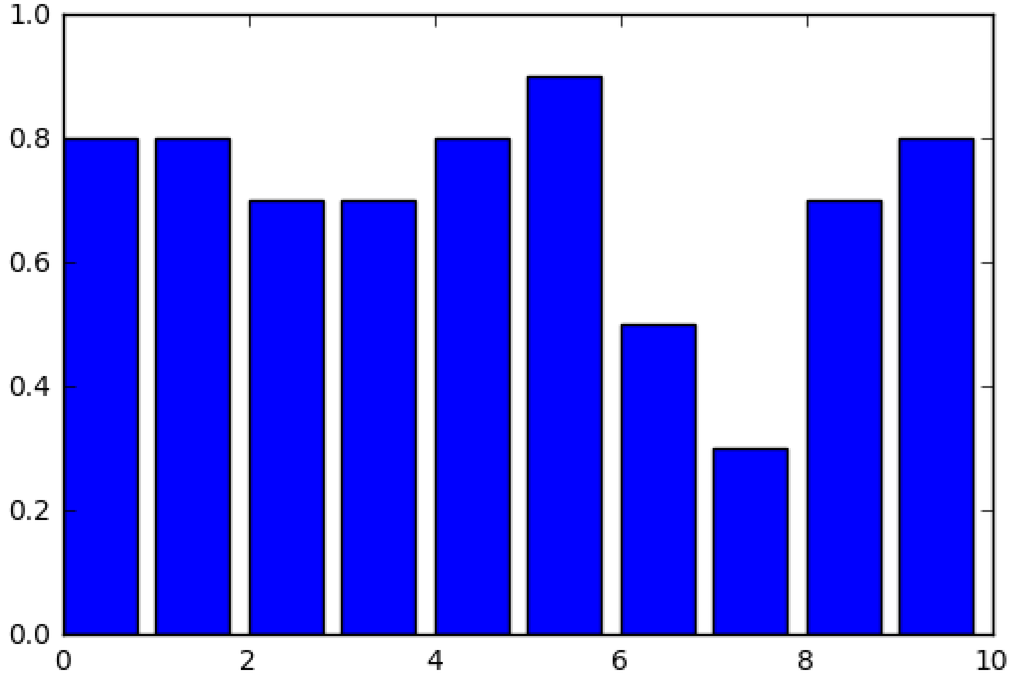
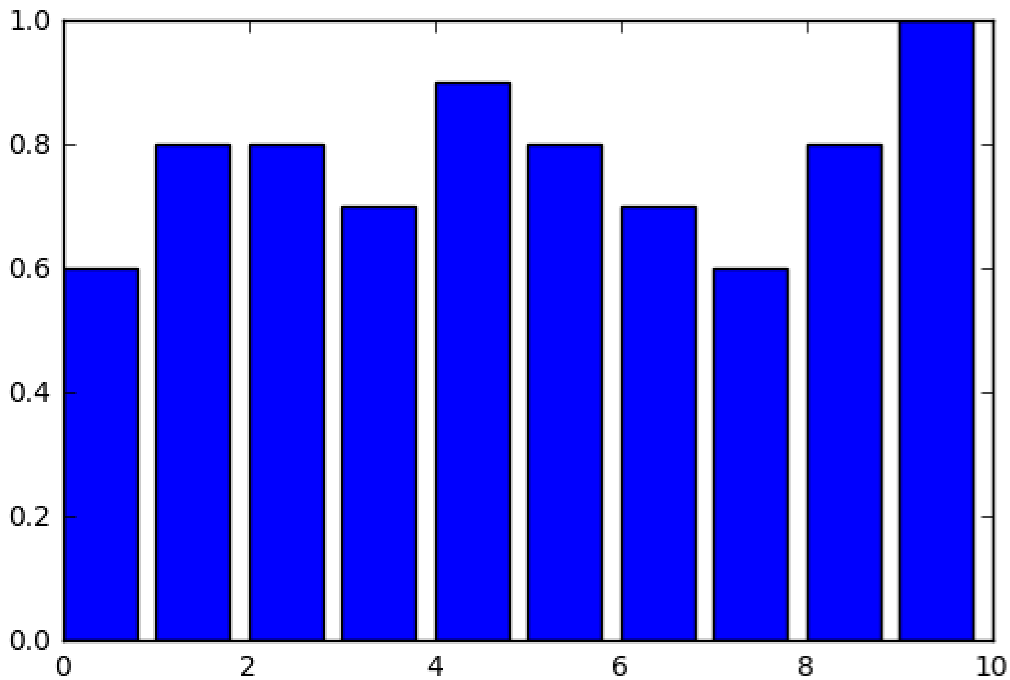


Figure 16: Accuarcy Across Videos With Middle Means and STDs



In terms of feature importance (table 7), now upper gamma waves are more important, and theta waves have decreased importance.

Table 7

	Importance
Mid Alpha 2 Mean	0.225020
Mid Gamma2 STD	0.141188
Mid Gamma2 Mean	0.118993
Mid Theta Mean	0.103224
Mid Theta STD	0.069433
Mid Beta 1 STD	0.067186
Mid Beta 1 Mean	0.043637
Mid Gamma1 STD	0.040002
Mid Alpha 1 STD	0.034248
Mid Alpha 1 Mean	0.032340
Mid Delta Mean	0.032024
Mid Beta 2 STD	0.029475
Mid Alpha 2 STD	0.026793
Mid Gamma1 Mean	0.021867

Mid Beta 2 Mean	0.009855
Mid Delta STD	0.004714

We can also investigate adjusting hyperparameters for this model, in particular the number of estimators, max depth, and learning rate. We find that a smaller number of estimators performs better, with 50 estimators getting an accuracy of 0.812500, and a brier score of 0.165893. This suggests the model is overfitting with the higher number of estimators. So we will keep the number of estimators at 50. However with both the max depth and learning rate, the default values outperform any adjustments we make.

Results

Model Evaluation and Validation¶

With the new parameter we can again check to see what features it's looking at in Table 8. The biggest change from before is an increased importance on theta waves. They remain less important than upper alpha waves, but the combined importance of their mean and standard deviation are very close to those of alpha 2 waves.

Table 8

	Importance
Mid Alpha 2 Mean	0.256545
Mid Theta Mean	0.160155
Mid Gamma2 STD	0.151073
Mid Theta STD	0.087094
Mid Gamma2 Mean	0.081805
Mid Beta 1 STD	0.079264
Mid Gamma1 STD	0.057449
Mid Alpha 1 Mean	0.031012
Mid Beta 2 STD	0.023794

Mid Beta 1 Mean	0.018960
Mid Alpha 2 STD	0.013502
Mid Gamma1 Mean	0.013373
Mid Beta 2 Mean	0.010215
Mid Delta Mean	0.010031
Mid Alpha 1 STD	0.005424
Mid Delta STD	0.000305

If we look at how well it generalizes across students (figure 17) we see it does almost as well with a mean accuracy of 68%. It performs the same across videos (figure 18) with a mean accuracy of 77%.

Figure 17: Accuracy Across Students

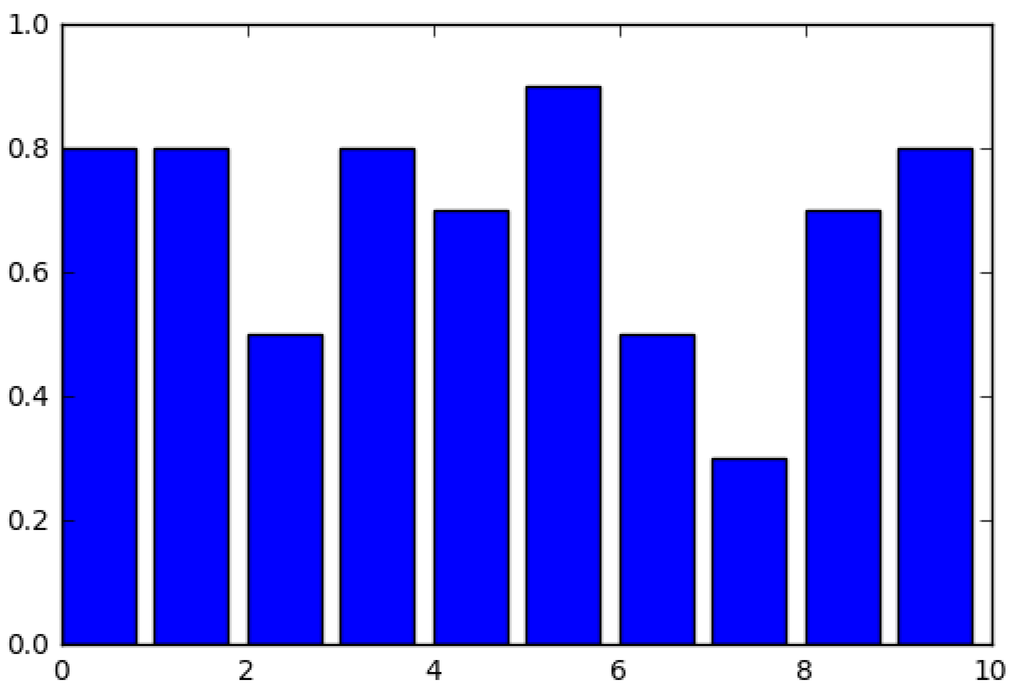
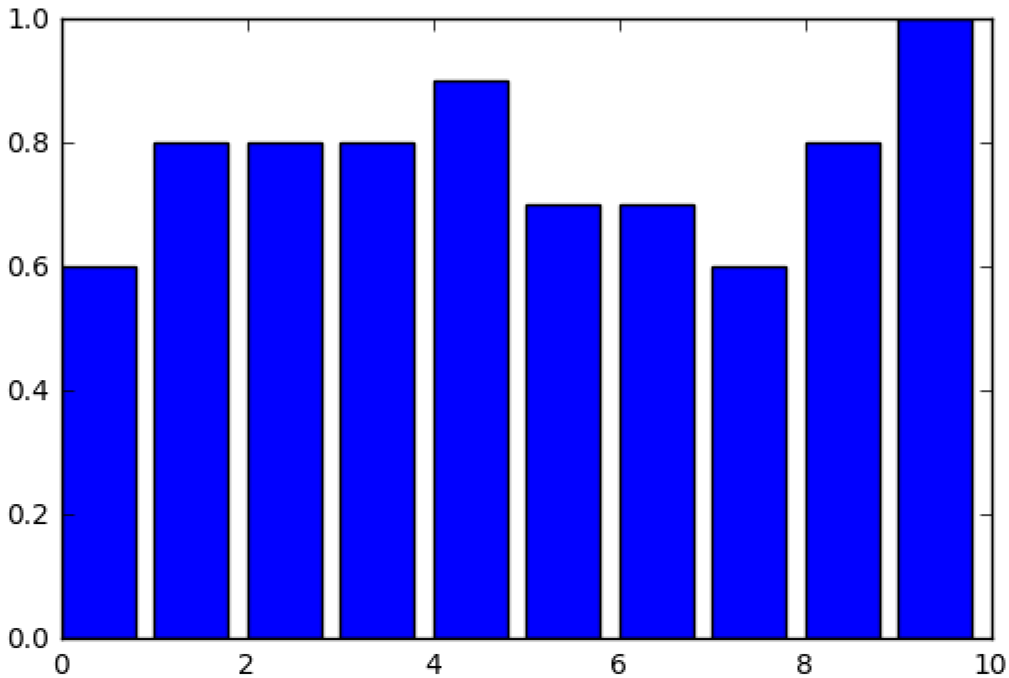


Figure 18: Accuarcy Across Videos



Justification

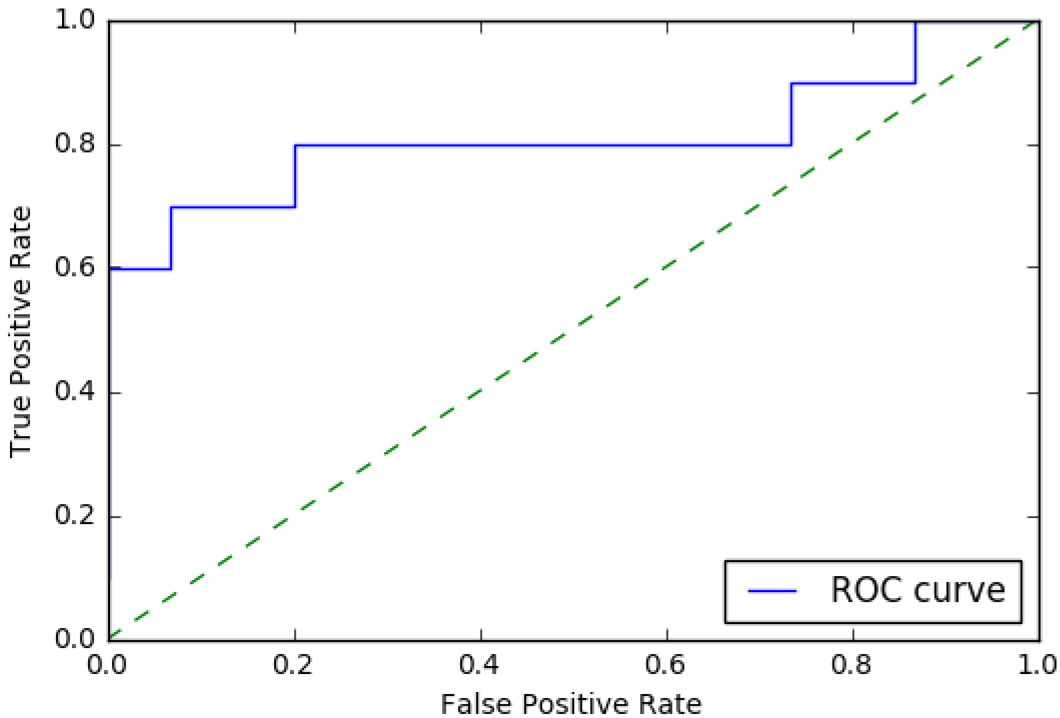
We are getting an overall accuracy of over 80%, which is significantly better than the 65% percent reported by the researchers as a "decent model" or the 70% I identified as my goal. This model seems to do very well across different videos, though it remains inconsistent across different students. This is not too unexpected, as different brains do work differently. However it means that any use of neural data for confusion will need to do some calibration for a new student.

Conclusion

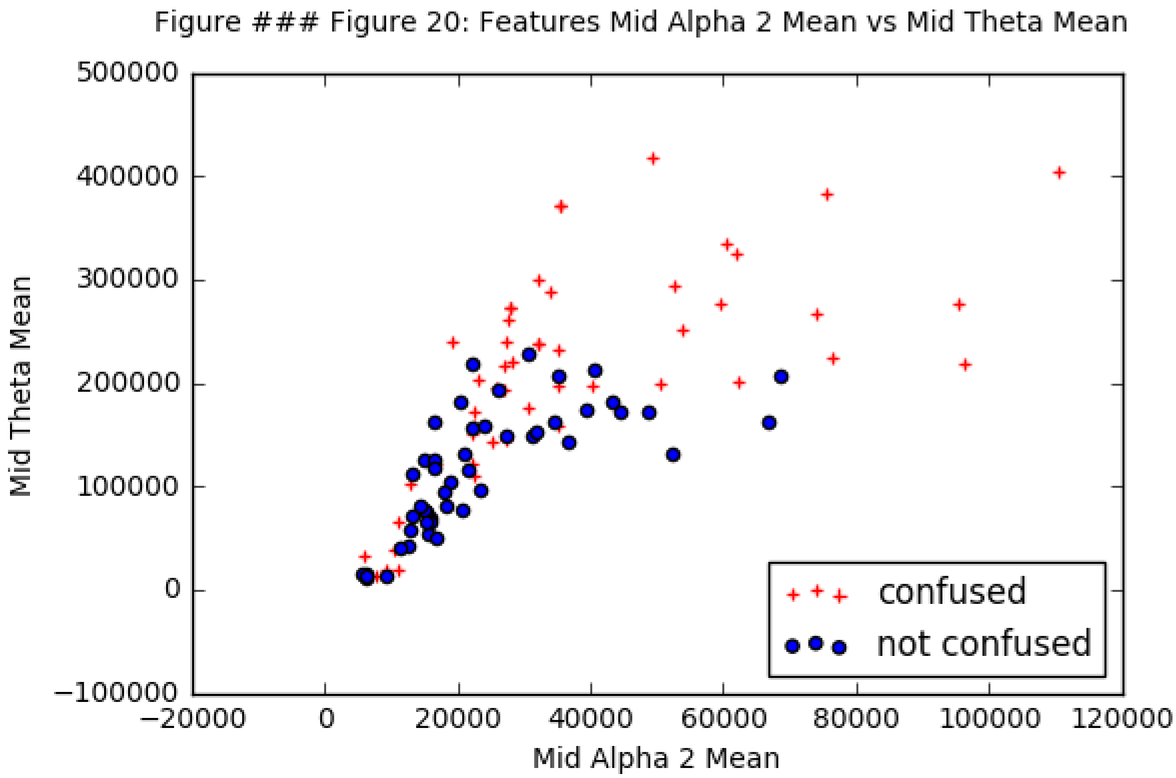
Visualization

To get a better idea of how well this model performs, we can look at the ROC curve it generates in Figure 19.

Figure 19: ROC for optimized classifier



Since the model is taking particular interest in the Alpha 2 and Theta waves. We can plot them to get a visualization of how they appear in different sessions in Figure 20. Here we can get a good idea as to why the model is treating both of these features with high importance. They are roughly correlated, with higher values of both being associated with confusion.



Reflection

For this project I was able to make a well performing model for determining if a student was confused by a video based solely on their brain waves as detected by an EEG. A Gradient Boosting Ensemble model, looking at both the mean and standard deviations of the amplitudes of brain waves within 8 different frequency ranges during the middle of the video, is able to achieve an accuracy of around 80%. By looking at the model we can also get an idea of what features are important in determining if a student is confused. Previous research had suggested the importance of theta waves, but this model is also considering upper alpha and upper gamma waves.

Improvement

As successful as this model is, it is based on very limited data. We only have 10 different students and 10 different videos resulting in 100 different data points. A wider selection of data would make for a more robust dataset.

Furthermore we are grading the confusion levels with a very coarse grained and subjective evaluation. Different students will have different internal metrics for what "confusion" means, and we don't really get to take in consideration the level of confusion a student has. It may be some of the false predictions involved students who were on the threshold of being confused or not. A continuous confusion level modeled with a regression model may be able to capture some of that. And if the confusion level were captured in a more objective fashion (such as giving the student a short quiz on the material) we may be able to be more confident in the results.