Exploration of classification methods for interpreting auditory attention in BCI EEG data

Siena Guerrero (Harvey Mudd College)

in collaboration with Rishov Chatterjee (Pitzer College), Teresa Ibarra (Harvey Mudd College),

and SiKe Wang (Claremont-McKenna College)

Acknowledgements

The authors would like to thank M.S. Treder, H. Purwins, D. Miklodv, I. Sturm, and B. Blankertz for their original work on this subject and, specifically, for the EEG data that made our analysis possible. The authors would also like to thank Professor Michael Spezio at Scripps College for his helpful guidance, comments, and support on this project.

Abstract

How close is the field of neuroscience to reading minds? With current BCI methods, as shown by Treder et al., it is possible to read a subject's EEG neural signal and classify, with reasonable accuracy, if that subject was paying attention. In our project, we expanded upon the work of Treder et al. in order to explore the efficacy of using alternative machine learning methods when analyzing the original experiment's neural signals. To accomplish this goal, we classified attention using 3 different classifiers from the original paper: Random Forest, Logistic Regression, and Neural Nets. Using our optimized classifier, we were able to improve upon the original experiment's performance, but we could not make any definite conclusions about the underlying neural code of attention. Given that our feature importance vectors from our experimental classifiers did not share many large regions of similarity, our results imply that our classifiers were not representative of the brain's actual operation or that one method is correct while the others are not. In conclusion, we can make accurate predictions in classifying attention, but the science still has a long way to go before reading minds from EEG signals is a viable outcome.

Keywords: Machine Learning, signal analysis, brain-computer interface (BCI), electroencephalography (EEG), feature extraction, random forest, logistic regression, neural nets

Introduction

One emerging area of interest within the field of neuroscience has been the application of machine learning to assist in the development of brain-computer-interfaces. As computers have become faster and better able to process neural signals, the field of Brain-Computer-Interfaces has expanded; in 2004, Schalk et al. proposed BCI2000 as a method for implementing BCI systems, and a decade later Treder et al. used a similar BCI system to explore the classification of auditory attention with BCI EEG data. By performing experimentation that collects and analyzes neural signals, scientists can gain a better idea behind the underlying neural code. Treder et al. performed such experiment in 2014 while attempting to classify auditory attention in the brain, producing a dataset where attention can be inferred from its corresponding EEG neural signals (Treder, 2014). In the original Treder et al. Paper, the dataset was analyzed only using a single classifier with linear discriminant analysis. The eleven subjects were instructed by the researchers to listen to the music being played and to attend to particular instruments being played. During this process, the researchers captured EEG data from the subject using 64 electrodes for three time epochs.

In an attempt to improve upon the work of the original authors, we have analyzed the original dataset with additional machine learning techniques by using three additional classifiers, and we have investigated classifying auditory attention by measuring ERPs. In order to best understand these spatiotemporal features, we performed analysis using the same 3 time epochs that the original authors determined from their dataset's event-related potentials.

In order to research different techniques that would be useful in analysis, we explored different machine learning methods in *Pattern Recognition and Machine Learning* (Bishop,

2016). Using the information described in this textbook, we built several models designed to be trained and subsequently tested using different classifying methods, including logistic regression with L2 regularization, neural networks, and random forests.

In preparation for analysis, the training set was prepared by our team to consist of a randomly chosen set of seven subjects to perform training and validation. The remaining four sets of subjects were used to test our classifier for performance, using each individual subject as a separate test case. During cross-subject testing, we wanted to see if there was similarity between our chosen three classification methods, their feature importance vectors, and shared areas of importance within the brain. Current studies on neuroimaging indicate that activity in the frontal and parietal regions of the brain likely corresponds to attention (Seydell-Greenwald, 2014). So, we would predict if models are performing correctly, that the feature importance matrices that arise from these neural signals would be clustered in these particular areas across subjects.

One important note is that, when it came to cross-subject classification, neural nets did not seem to suffer from the sort of over-classification issues exhibited by random forests (The Random Forest Algorithm, 2018). If the issue of unbalanced data was solved during our project, we may have seen different performance metrics for random forests that could have made it better than neural nets.

Overall, our ultimate goal in this project is to show a similarity between neural code and auditory attention in the brain in order to gain a better understanding of how thoughts, like attention, can be transformed into digital processes with the aid of a BCI system (Thorpe, 2005). To aid in this understanding, we chose to use machine learning practices whose feature

importance vectors might provide confirmation of neural patterns that we can ascribe to auditory attention in the human brain. While we were able to improve upon the original authors' classifier performance, we were unable to find a compelling region of similarity for the underlying neural code, which we be expanded upon later in this report.

Methods

Research Design

Before performing full analysis, we decided to investigate several different classifiers in order to get a gauge on how successful some of them might be at interpreting our data and detecting patterns in order to ultimately give better accuracy on our testing sets. Some of the methods we explored were linear discriminant analysis for linear separation, logistic regression and neural networks for a probabilistic perspective, and random forest for nonlinear separation. Once we had conducted preliminary testing, we chose the three models that had performed the best (in terms of accuracy) to explore further with our full dataset and specified best model parameters. The three models that performed the best were neural nets, random forest, and logistic regression. In order to prevent overfitting our datasets, we used cross-validation and investigated bias-variance tradeoffs. Unfortunately, we neglected to account for unbalanced datasets, which may have contributed to poor perform with Random Forest during individual subject testing.

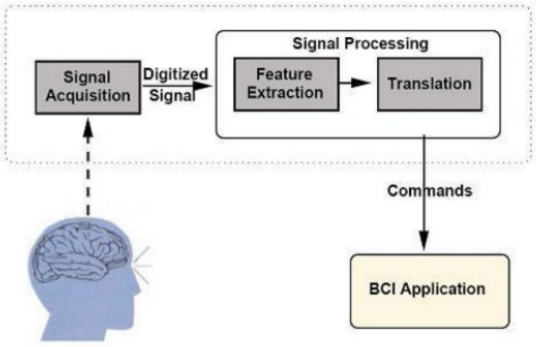


Image caption: Basic design and operation of any BCI system. Signals from the brain are acquired by electrodes on the scalp, the cortical surface, or from within the brain and are processed to extract specific signal features (Thrope, 2015)

Our goal with this dataset is to classify cross-subject attention. First, we split our dataset into two, with a test and training set. The training set was prepared to consist of a randomly chosen set of seven subjects to perform training and validation. The remaining four sets of subjects were used to test our classifier for performance, using each individual subject as a separate test case. In essence, we concatenated the 7 subjects' features and labeled data as one super set that training was conducted with according the our three models. Finally, the L2 regularization, neural network, and random forest training models were tested for performance with the remaining four sets of subjects.

In order to optimize cross-validated accuracy, we found the best model parameters using randomized search with each of the three models. Once these specific hyperparameters were determined, we could generate confusion matrices, ROC curves, and accuracy that can aid in interpreting classification performance for each individual model.

Confusion matrices are tables that tell us certain metrics like our error rate, sensitivity, false positive rate, true negative rate, specificity, precision, and prevalence. The ROC curves give information about model sensitivity across different threshold settings.

Accuracy gives some insight into interpreting classification performance and particularly in establishing a link with cross-subjects, but it is very important to remember that it does not tell us much about our original question of how the underlying neural code relates to attention in humans. In order to better understand this part of our problem, we must examine the feature importance vector for each model, as this information can then be used to map the importance of the 63 electrodes attached to the subject If there is similarity in the feature importance vectors among the subjects, then we can infer which regions of the brain may correspond to auditory attention.

Logistic Regression . We decided to use logistic regression with L2 regularization for our first model.

Logistic regression is a classification algorithm that is used to assign observations to a set of classes, which in this experiment is attended vs. non-attended. Logistic regression takes its output and transforms it with the logistic sigmoid function into a probability value. This value is then used to assign the observation to some discrete class. Because we only have two possibilities for classification, attended vs. non-attended, there are two discrete classes.

We used logistic regression with L2 regularization, which is useful in dealing with data that exhibits multicollinearity, or when one variable can be predicted by another. Using this particular for of logistic regression will help reduce the standard error for our data, aiding in classification.

Neural Nets. We decided to use logistic neural nets for our second model.

Neural nets is a linear classification method that functions through supervised learning. It functions by performing predictions using layers of that combines a set of weights with a feature

vector for classification. Overall, the network would be comprised of an input layer, a hidden layer, and an output layer, where each node acts as a neuron that has a nonlinear activation function.

Similar to Logistic Regression, we only have two possibilities for classification, attended vs. non-attended, our neural net would have one output node corresponding to the generated classification.

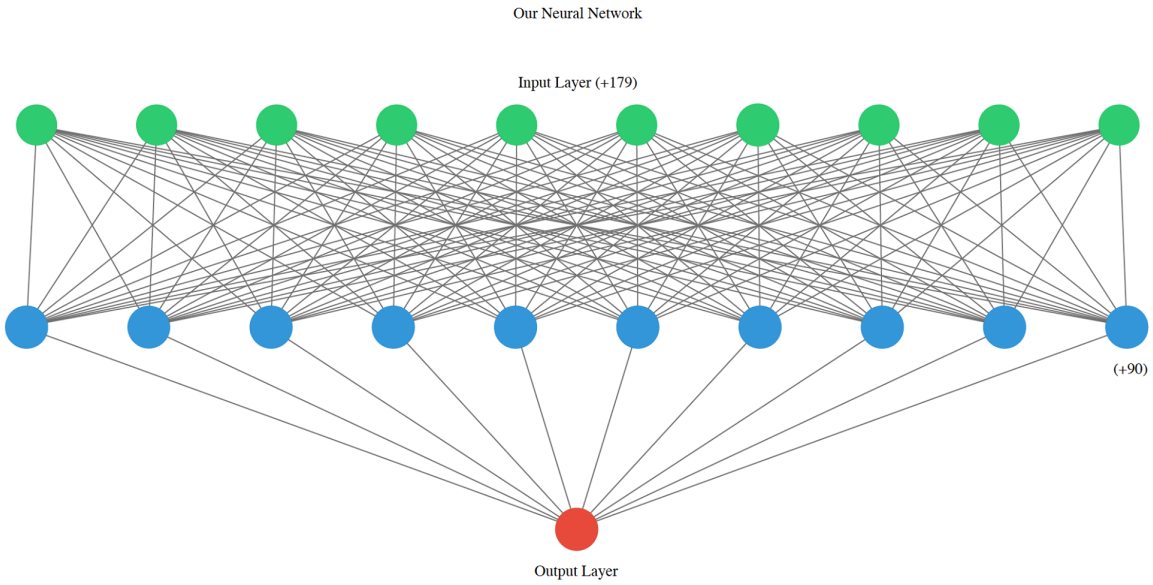


Image caption: This figure represents the multiple layers of our neural network: an input layer with 189 nodes (the number of features), a hidden layer, and an output layer with one node. (Courtesy of Rishov Chatterjee)

Random Forests. We decided to use random forests for our third model.

Random forests constructs multiple decision trees and then outputs the class that is the mode of the trees' classes. In our case, we constructed 1000 decision trees. A decision tree is a structure in which each node represents a test on some attribute of a given observation. Each

branch in the decision tree represents the outcome of the test and each leaf node represents a class label. The class labels, as mentioned before, are attended and unattended.

We chose different methods for splitting random trees: Gini and Entropy. In essence, these methods are the criteria for determining branch layouts in generated decision trees by acting as methods for determining "splits" in the data for deeper trees. For the majority of tests, Gini was determined to correspond to better performance.

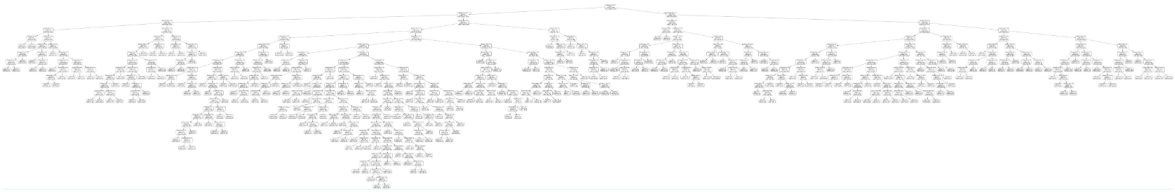


Image caption: Visualization of an example of a decision tree in our random forest model with Gini Index criterion (Courtesy of Siena Guerrero)

Existing Dataset

According to the information provided by Treder et al., the dataset is composed of trials from eleven different participants, of whom seven were males and four were females. The subjects were aged 21–50 years with a mean age of 28, and all but one were right-handed. For our experiments, we used the dataset in order to perform both intra-participant and inter-participant analysis, but in the original paper, the authors only used the data for within-subject classification. The paper notes that "each trial started with a visual cue indicating the to-be-attended instrument. Then, the standard stimulus and the deviant stimulus of that particular instrument were played." After this point, the participants were instructed to count the number of deviants in the attended instrument, as this method was intended to ensure that the subjects showed clear neural evidence of attention.

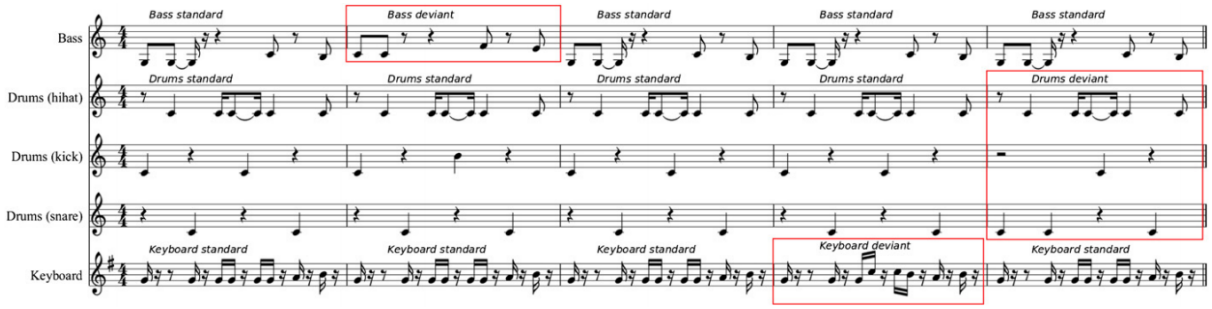


Image caption: Score sheet illustrating the multi-streamed musical oddball paradigm for the Synth-Pop stimulus. (Treder, 2014)

The original experimenters used an electrode system with 64 electrodes that enabled capture of 1000 Hz EEG neural signals. The first electrode, which corresponded to subject blinking, was not used in the feature matrix. In addition to the neural signals, the dataset also includes timing information about the events and deviants during EEG capture.

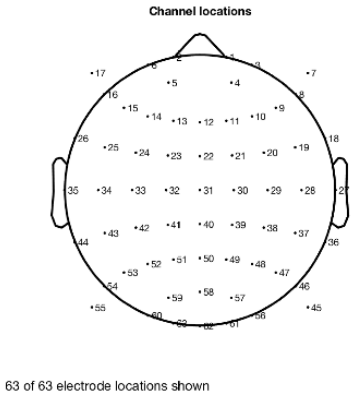


Image caption: Locations of 63 electrodes used in this experiment as sources of EEG signals for further classification tasks. (Courtesy of Michael Spezio)

The EEG data were pre-processed with the BBCI Matlab toolbox where the signals were "downsampled to 250 Hz and low-pass filtered using a Chebyshev filter," after which point the data could then be split into time epochs. Using a minimax criterion, the original authors rejected

artifacts, with the stipulation that they were preserved in the test-set for classification purposes. The original researchers decided on a range from −200 ms prestimulus to 1200 ms poststimulus for the time epoch for each deviant. In order to classify the neural signals, the classifier is only used to identify attended and unattended deviants. As such, it does not consider standard stimuli, and so the time epochs were specifically chosen to correspond with peaks in the point-biserial correlation coefficient sgn r^2 between attended and unattended deviants after the stimulus. Once these three time intervals were determined, the researchers averaged the voltages in these time periods and had this calculation recorded as one of the electrode and three epochs' features.

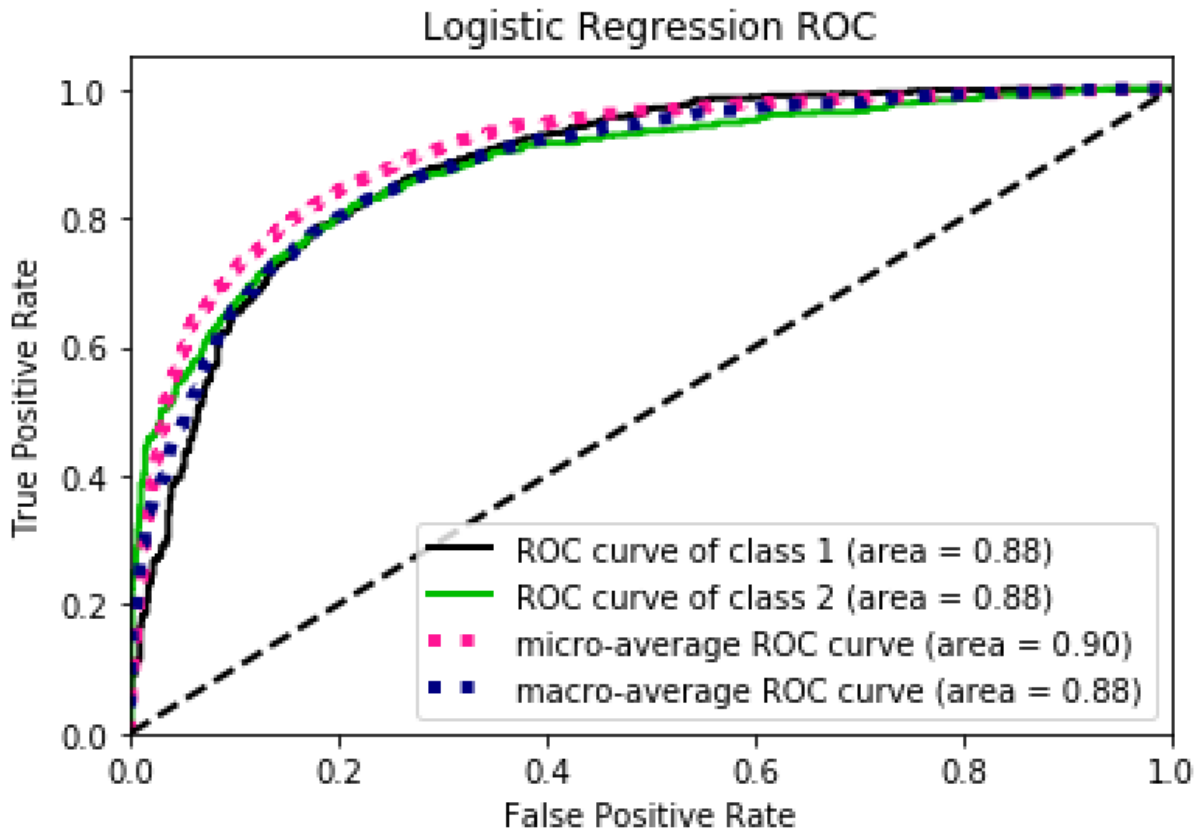
Signed Pointwise Biserial Correlation

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}}$$

Image caption: The point-biserial correlation coefficient sgn r2 that leads to the selection of poststimulus intervals of attended deviants and non-attended deviants. (Courtesy of Michael Spezio)

Results

Logistic Regression with L2 Regularization:


Logistic Regression ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 1 (area = 0.88)
ROC curve of class 2 (area = 0.88)
micro-average ROC curve (area = 0.90)
macro-average ROC curve (area = 0.88)

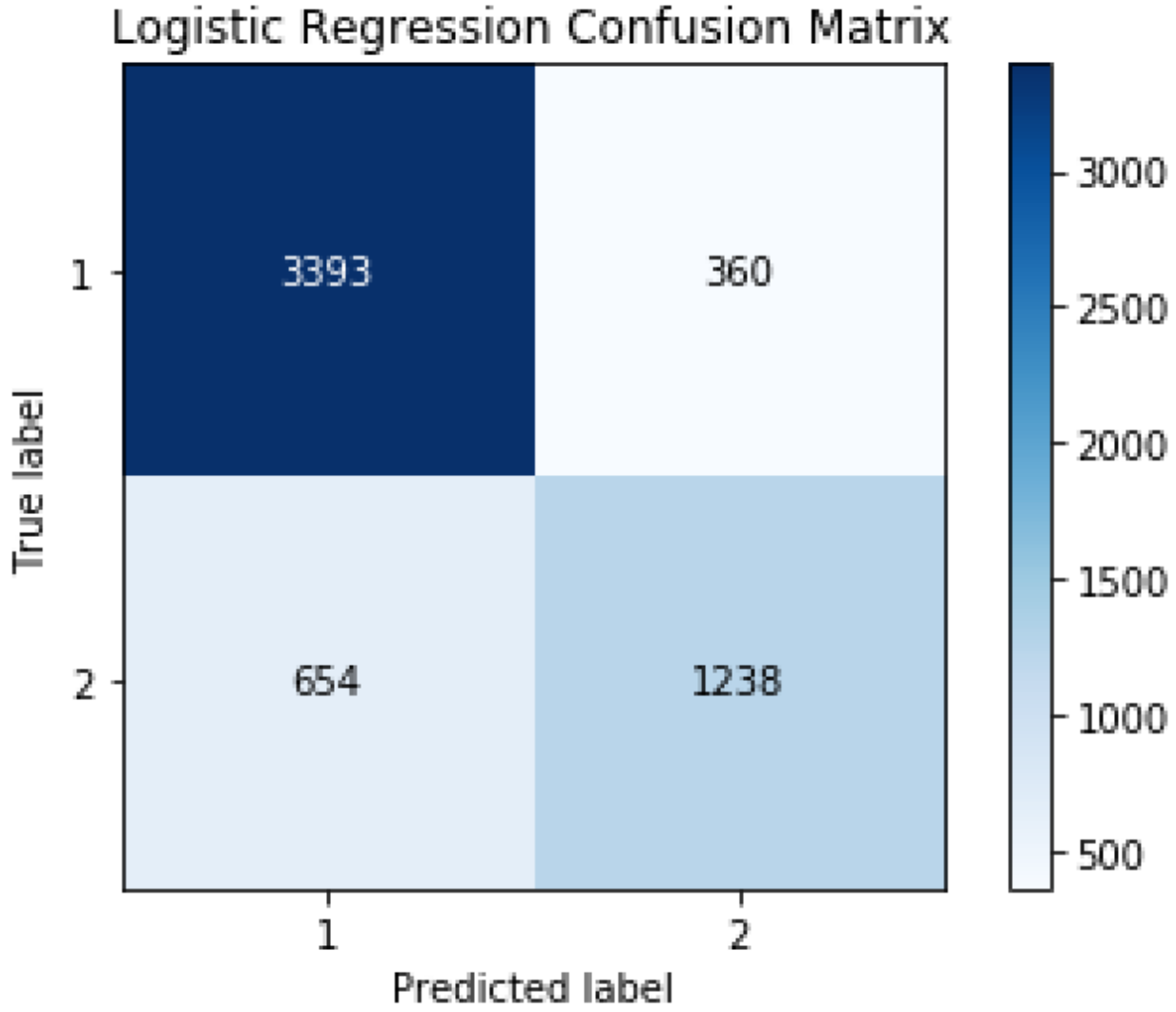
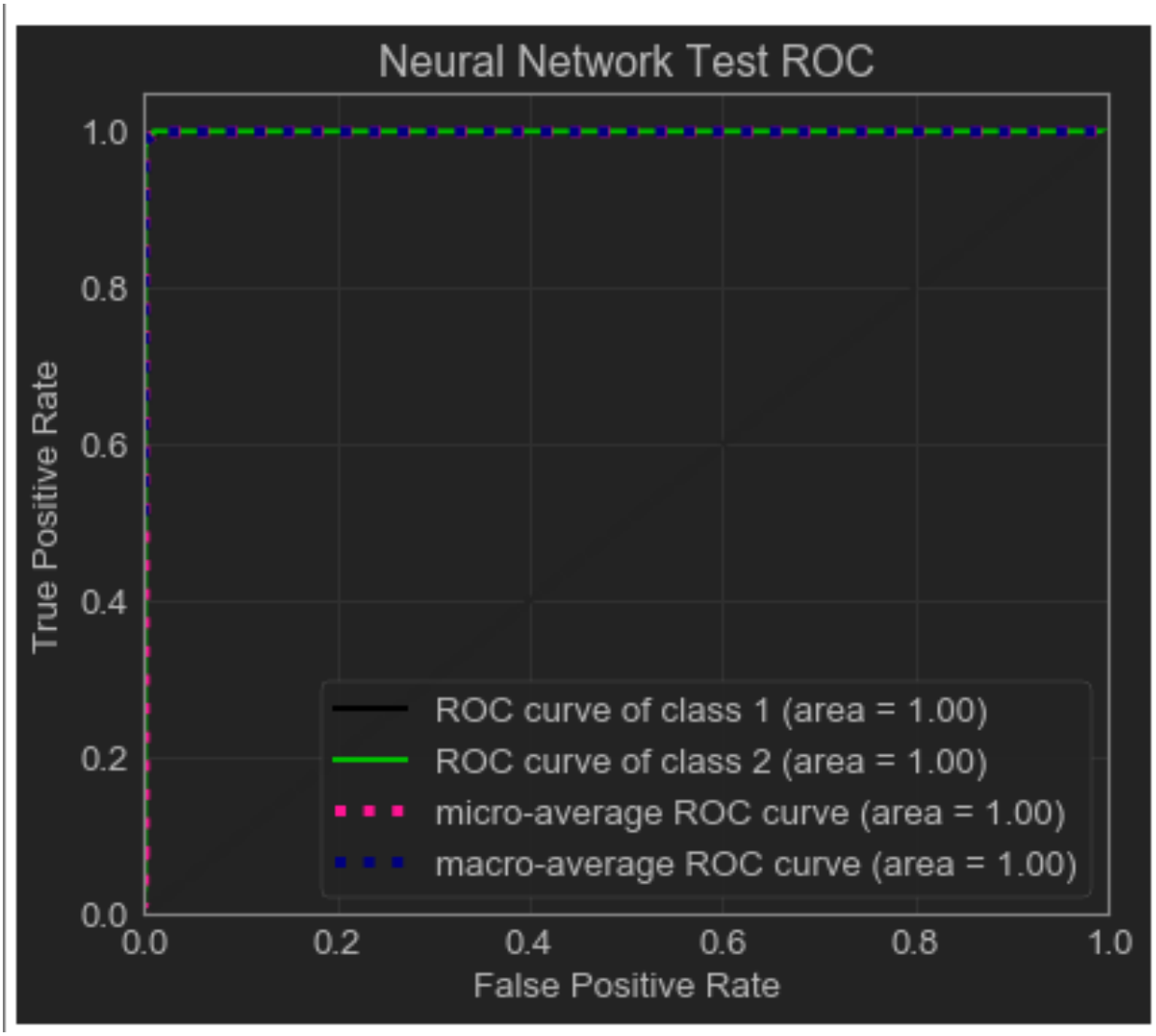


Image caption: Figures are the ROC curve and confusion matrix respectively for subject aap with logistic regression with L2 regularization.

Neural Networks:


Neural Network Test ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 1 (area = 1.00)
ROC curve of class 2 (area = 1.00)
micro-average ROC curve (area = 1.00)
macro-average ROC curve (area = 1.00)

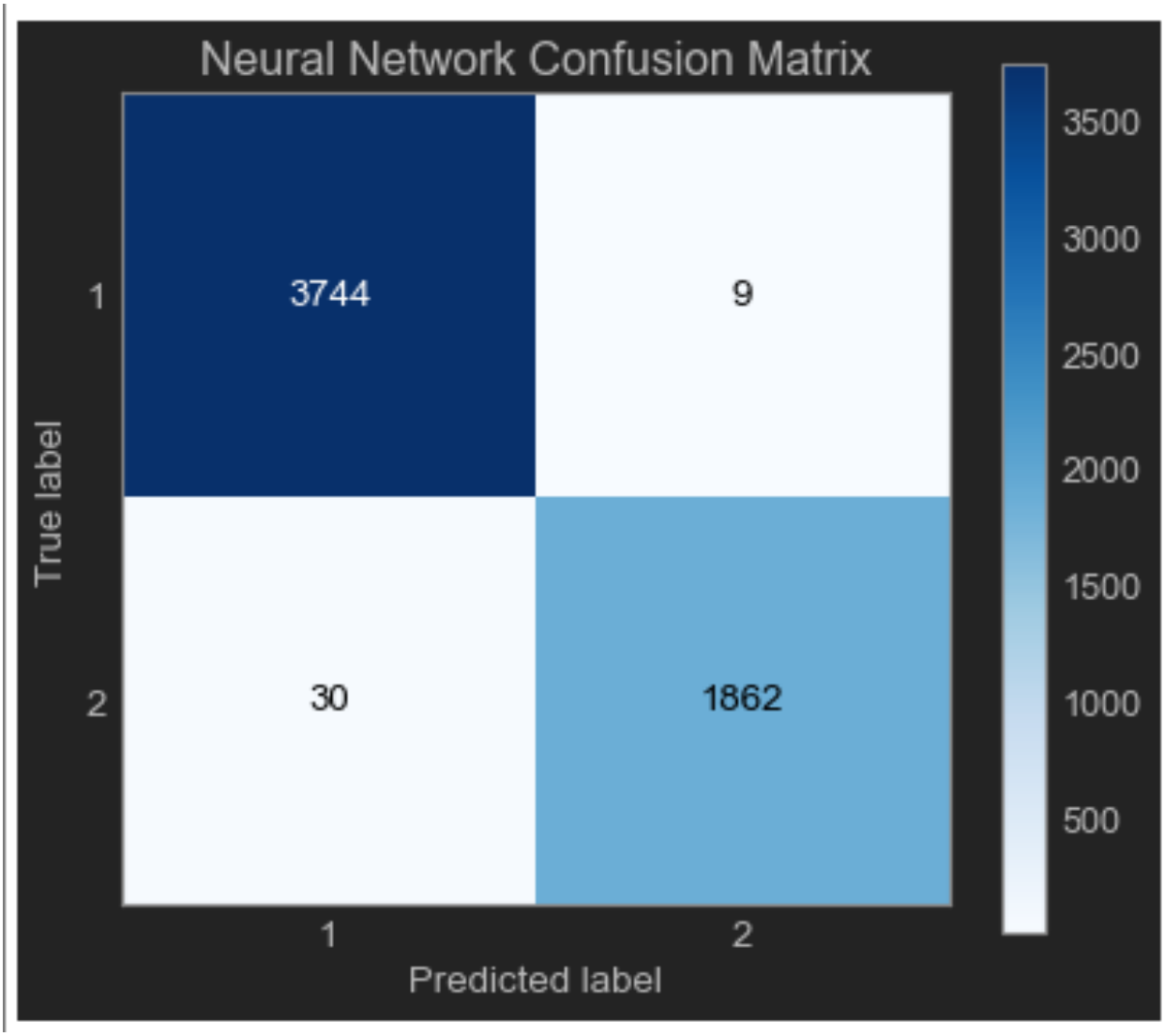
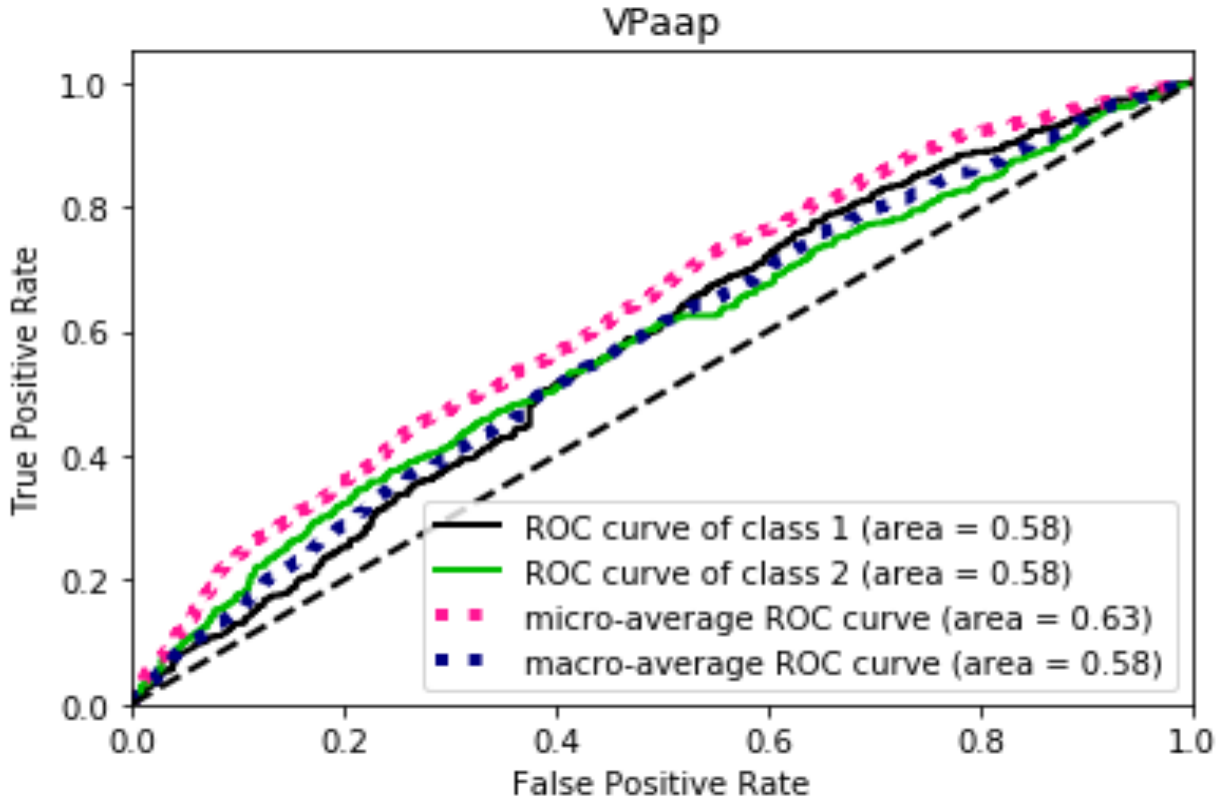


Image caption: Figures are the ROC curve and confusion matrix respectively for subject aap with a neural network. The accuracy score reached 99.31%.

A three-layer perceptron did very good classification for all subjects, with accuracy scores of 99.31% across all participants. This shows that the MLP algorithm was able to smooth out individual differences and generate across subjects relatively well, even when it encounters an imbalanced dataset.

Random Forests:


VPaap
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class 1 (area = 0.58)
ROC curve of class 2 (area = 0.58)
micro-average ROC curve (area = 0.63)
macro-average ROC curve (area = 0.58)

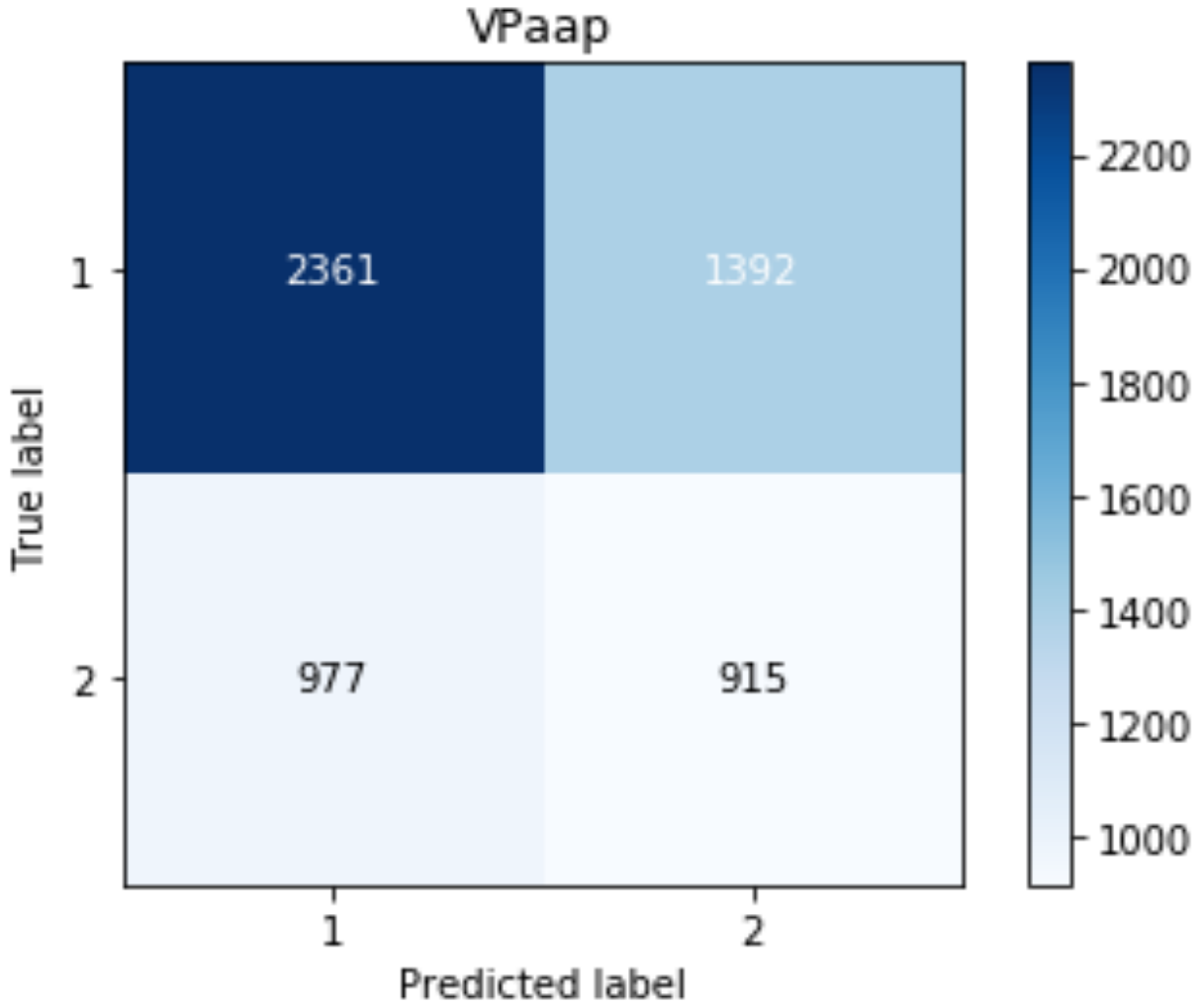
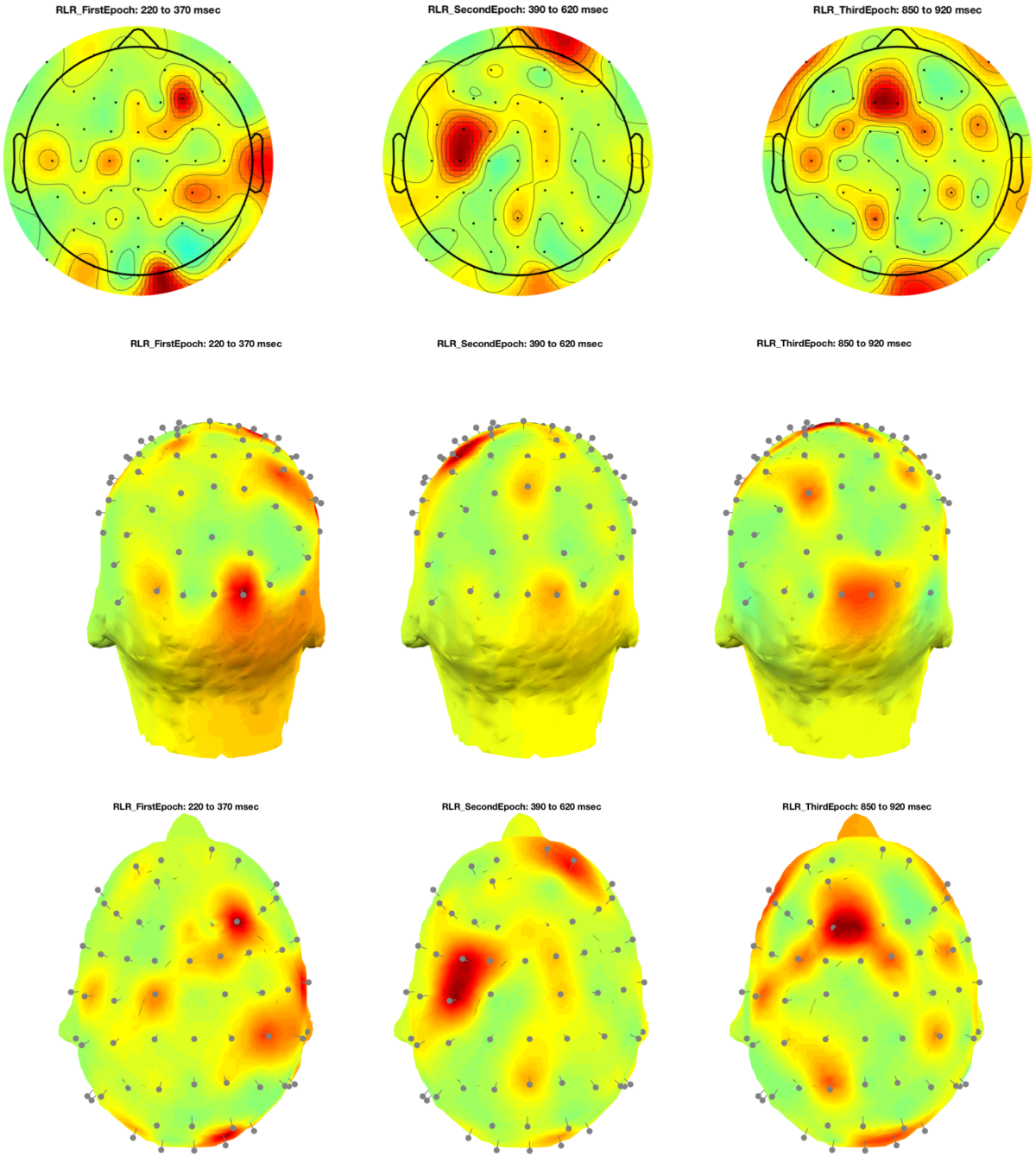


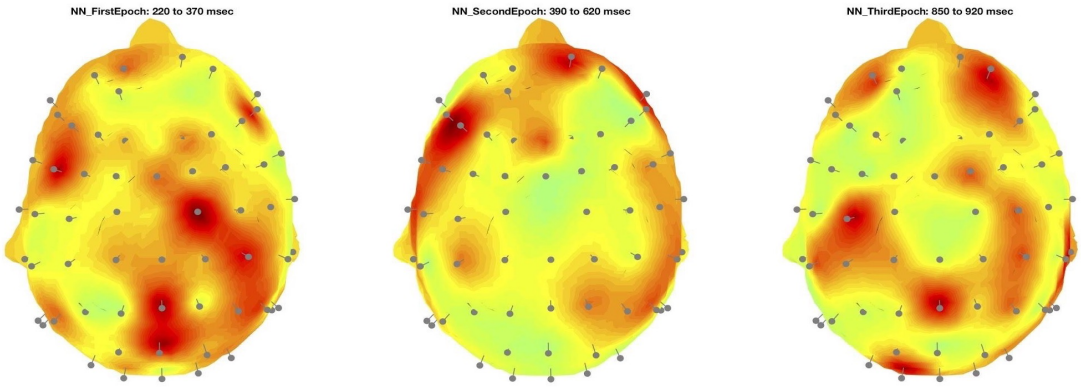
Image caption: Figures are the ROC curve and confusion matrix respectively for subject aap with a random forest. The accuracy score reached 58.03%, likely due to class imbalance.

Overall, all three models indicate that the results across all subjects were very similar, with results above those predicted by random classification.The only connecting major information that some of the confusion matrices show is evidence of bias due to the large number of true positives and false positives in combination with a low amount of false negatives and true negatives. This sort of result could indicate that there is likely an issue of class imbalance, .meaning that our models over-predicts if a subject is paying attention, to the point of misclassifying subjects who are not paying attention. One important note is that neural nets, which performed the best overall, also did not seem to suffer from the sort of over-classification issues exhibited by random forests (The Random Forest Algorithm, 2018).

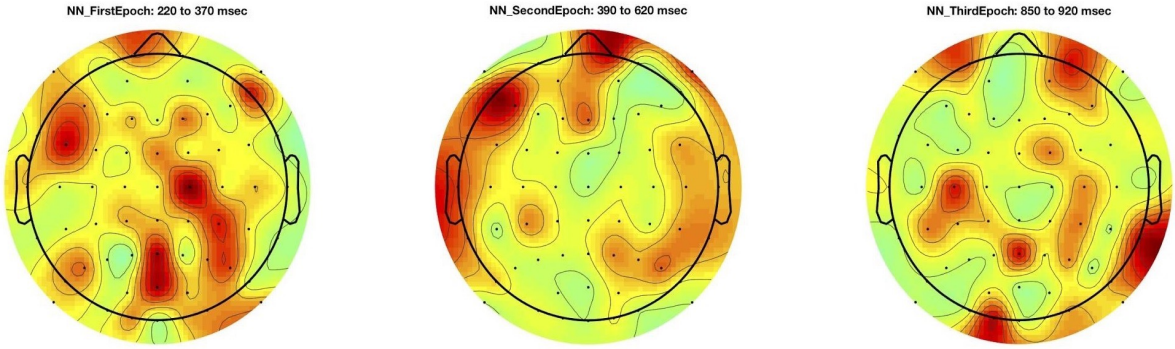
Feature Importance Vectors for Three Models (Generated by Michael Spezio):

Logistic Regression

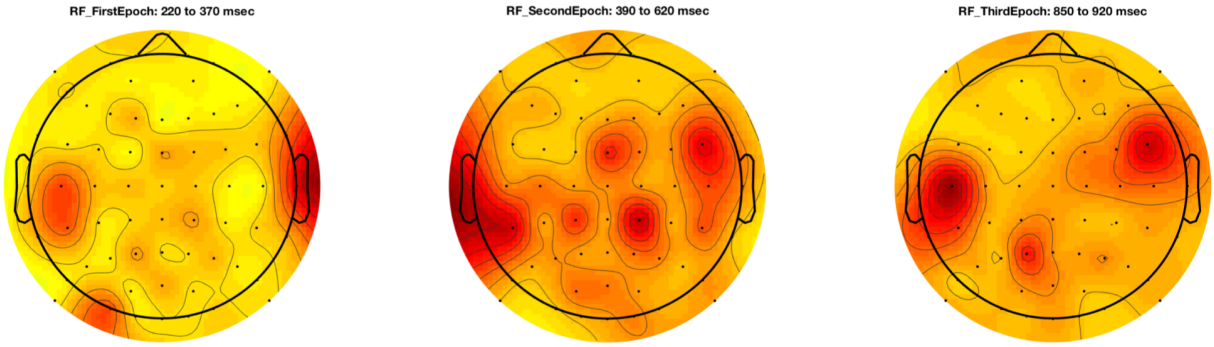




Neural Networks (median values)



Random Forest (Gini classifier)



We would expect the features extracted from all three models to be concentrated primarily in the frontal and parietal regions of the brain in a manner that is consistent across classification methods. However, our generated images from the models' feature importance vectors did not appear to show significant clustering that is reasonably shared among the three models. We can

see some expected activation in the regions for the first and last epoch with logistic regression and neural nets, but overall neural nets shows a wider spread of brain clustering. In general, Gini and Entropy testing for random forests shows clustering in the same regions, but the weights for the significant brain regions is much stronger than what can be seen with the logistic regression and neural nets models. These results make some since given that the model weights are all approximated from different mathematical functions that could result in different outcomes.

Randomized Search Results

	Cross-validated accuracy	**Best model parameters**
Logistic regression with L2	82.09%	penalty=L2, intercept_scaling=1.75, fit_intercept=True
Neural nets	99.34%	activation=relu, solver=adam, shuffle=False, learning rate=adaptive, alpha=1e-08, hidden layer size=100
Random forests	99.3%	n_estimators=800, max_depth=None, criterion=gini, bootstrap=True

Accuracy Table

	Logistic Regression	**Random Forests**	**Neural Networks**
aas	82.04%	58.03%	99.31%
gcc	82.04%	58.03%	99.31%
aap	82.04%	58.03%	99.31%
aan	82.04%	58.04%	99.31%

Discussion

Conclusions

We came to three main conclusions upon completion of our project.

1. We were unable to concretely determine which time epoch out of the three was the most important to consider when classifying attention.
2. We were unable to concretely determine a similarity in neural code among subjects, or more broadly, how/why auditory attention is shared across humans.
3. Our three methods for classification, logistic regression, neural nets, and random forests, did not share prominents similarities in their feature importance vectors despite high training and testing performance.

Implications

We have two important questions to consider when analyzing our results: how well did our models perform in comparison to the original researchers' methods, and what insight can our resulting importance vectors give on the underlying neural code? In regards to the first question, in general, our models outperformed the researchers. Our neural net performed particularly strongly in classifying attended vs non-attended. However, the feature importance vectors that result from our models do not give a clear connection to areas of the brain among different subjects.

This result tells us that either the brain is not operating according the method performed by our models, or it is operating by the method of one of the models and not the other two. Therefore, there is little useful information that we can deduce from the neural code. Aside from the disconnect between classifier methods, it is also important to note that it is difficult to generalize what constitutes a "successful" model because they are not guaranteed to return the same results.

Future Work

In the future, the work on this project could be expanded upon through further exploration of other critical variables in the experiment like instrument-specific classification, We did not

have time to test our models on our dataset focusing on what instrument the subject attends to, but the original authors have shown that such classification is possible. Also, additional work can be performed to intra- and inter-subject classification. Again, due to time, we were not able to run as many training/test sets as we would have liked in order to get a more clear understanding of the areas of similarity between subjects. Our experimentation also suffered due to a lack of accounting for unbalanced data among the subject datasets. Because there were more instances of attended overall, the classifier was more likely to skew towards false negatives when working with a classifier like Random Forest. This was not the case, however, with Neural Nets, indicating that such a technique may be loss prone to errors due to training on unbalanced data.

To account for these deficiencies, future researchers should conduct this experiment with a proper understanding of how balanced their data is as well as running this data using a wider assortment/quantity of random training sets. With these specifications in mind, future work may reveal feature importance vectors that given a clearer connection to the underlying neural code. In particular, more careful analysis could be performed on the specific time epochs attached to auditory attention, which may give insight into how long it takes the brain to properly register stimuli. The more optimized these measurements, the better BCI designers are able to design a system that seamlessly connects the brain and computer.

References

Artificial Neural Network. (n.d.). In Wikipedia. Retrieved May 02, 2009, from https://en.wikipedia.org/wiki/Artificial_neural_network

Bishop, Christopher M. (2006). Pattern recognition and machine learning. New York: Springer.

Donges, N. (2018, February 22). The Random Forest Algorithm. Retrieved from: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Logistic Regression. (n.d.). In Wikipedia. Retrieved May 02, 2009, from https://en.wikipedia.org/wiki/Logistic_regression

Schalk, G., Mcfarland, D., Hinterberger, T., Birbaumer, N., & Wolpaw, J. (2004). BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. IEEE Transactions on Biomedical Engineering, 51(6), 1034-1043.

Seydell-Greenwald, A., Greenberg, A.S., & Rauschecker, J.P. (2014). Are you listening? Brain activation associated with sustained nonspatial auditory attention in the presence and absence of stimulation. Human brain mapping, 35 5, 2233-52.

Thorpe, J., Oorschot, P. C., & Somayaji, A. (2005). Pass-thoughts. Authenticating with our Minds - NSPW 05.

Treder, M. S., Purwins, H., Miklody, D., Sturm, I., & Blankertz, B. (2014). Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. Journal of Neural Engineering,11(2), 026009.