Classifying Complex Cognitive Operations from fMRI Data

Caleb Jerinic-Brodeur, Owen Friend, & Lingwei Ouyang

SDS 384 - Scientific Machine Learning

Final Project - Spring 2024

Introduction

The ability to intentionally control the contents of our mind is a vital function of human cognition. Working memory allows individuals to flexibly guide goal-directed behavior by providing efficient access to useful, task relevant information. Due to its capacity limits (Cowan, 2001; Luck & Vogel, 2013), removing irrelevant or unwanted information from working memory allows for its efficient use (Lewis-Peacock et al., 2018).

Previous research has proposed that individuals can utilize cognitive control strategies to remove information from working memory (Banich et al., 2015; Kim et al., 2020; Lewis-Peacock et al., 2018). This work has identified three distinct strategies to remove a thought from working memory: suppress that specific thought, replace it with another thought, or clear the mind of all thoughts. Univariate analyses of functional magnetic resonance imaging (fMRI) data have revealed a set of cognitive control brain regions that are differentially recruited during each removal operation (Banich et al., 2015; Kim et al., 2020). Recent research has demonstrated that machine learning techniques (i.e., multivariate pattern analysis [MVPA]) can be used to analyze fMRI data (Haxby et al., 2014; Kriegeskorte et al., 2008; Lewis-Peacock & Norman, 2014). These techniques have been leveraged in prior research investigating working memory removal, which has found that the aforementioned cognitive control strategies are distinct in their neural correlates and can be classified from distributed patterns of neural activity (Kim et al., 2020).

The goal of this project was to classify cognitive control strategies used to manipulate information during a working memory removal experiment. This approach is similar to work done by Kim et al. (2020) but in a novel, unpublished set of data. While this is not an entirely novel approach, we reasoned that this was useful to pursue given the replication crisis in many

fields, including Psychology. While there are many similarities to the work of Kim et al. (2020), there are also important differences to consider. To start, the data used in this project is from a separate study that occurred multiple years after the Kim et al. (2020) study. Additionally, the current data was in a new set of participants on a different site, with different equipment and software. Finally, the current data only included three of the four previously mentioned cognitive control strategies (i.e., Maintain, Replace, Suppress).

Data

The data used in this project is a subset of the experiment from which it is derived (5 of 25 participants). This choice was made to reduce the computational cost of running classification algorithms. The raw blood oxygen level dependent (BOLD) data was masked with a whole-brain mask. In other words, we used features across the entire brain and did not restrict the data to certain regions of interest (ROIs). This resulted in a 2D array with a shape of (1098, 151424) [time-points x features] for each participant. We additionally loaded the corresponding study labels that indicated which cognitive control operation was being performed at a given time point, as well as what point of the trial corresponded to each time point (i.e., Encoding, Manipulation, ITI). All data was shifted forward by 5 TRs to account for hemodynamic lag. The label data was reduced to only include time points during the manipulation and ITI periods. These indices were then applied to the functional data to select only the data occurring at those desired time points.

Exploratory Analyses & Hypothesis Generation

In order to determine what type of classification model could best classify complex cognitive operations from fMRI data, we compared four classification approaches based on their classification accuracy and resulting Area Under Curve scores. In particular, we were interested

in how the complexity of the models we selected handled the inherent variability and non-linearity in our data. We hypothesized that, while simpler models like logistic regression or random forest with a shallow max depth could perform well on classification by avoiding overfitting, more complex models such as XGBoost or random forest models with deeper max depths may better capture the complexity of the neural signals supporting these complex cognitive operations. To that end, we implemented and compared four models: **L2 Logistic Regression**, **Random Forest with no max depth**, **Random Forest with shallow, medium, and deep max depths**, and **XGBoost**.

Modeling & Validation

To test whether our four models could classify each cognitive operation above chance, as well as quantify differences between each model, we implemented a between-subject cross-validation approach. To begin, each subject's neuroimaging data was processed using fMRIPrep, a common pre-processing approach for fMRI data including slice time correction, motivation correction, skull-stripping, and registration of each subject's neural data to template anatomy for direct comparison (Esteban et al., 2018). Next, time series data for each subject was extracted, resulting in 1098 timepoints and 151424 voxels (units of volume within the brain) for each subject. Feature selection was performed using SelectFpr in Python to reduce the number of features for each subject to 19347 based on the activation profile of each voxel.

Next, we employed our cross-validation approach, iteratively comparing subjects such that each classifier was trained on the preprocessed BOLD data from four of five subjects, and then tested on the fifth. For each model, accuracy scores and AUC were calculated to quantify model performance. For the logistic regression, we used L2 regularization and set hyperparameter *C* to 50. For the Random Forest model, we tested an unconstrained model (i.e.,

no pre-specified max depth) to a shallow model (max depth = 2), medium depth model (max depth = 10), and deep model (max depth = 20). After determining which max depth resulted in optimal performance of our Random Forest model, we implemented an XGBoost model with the same max depth. Finally, we used ANOVA and paired *t*-tests to directly compare each model's performance based on their resulting classification accuracy and AUC scores.

Results

L2 Logistic Regression: The results of our logistic regression model can be seen in **Figure 1**. Logistic regression has been used in prior work to classify cognitive control operations (Kim et al., 2020). The logistic regression model used an L2 regularization and the hyperparameter *C* was set to 50. The logistic regression model was successful in classifying all three operations above chance (**Figure 1A**). Additionally, the models for each operation performed above chance (**Figure 1B**). While all models performed above chance, there appears to be a greater amount of variability for Replace and Suppress. This may suggest that individuals engage distinct strategies, and thus distinct neural patterns, to complete these control strategies.

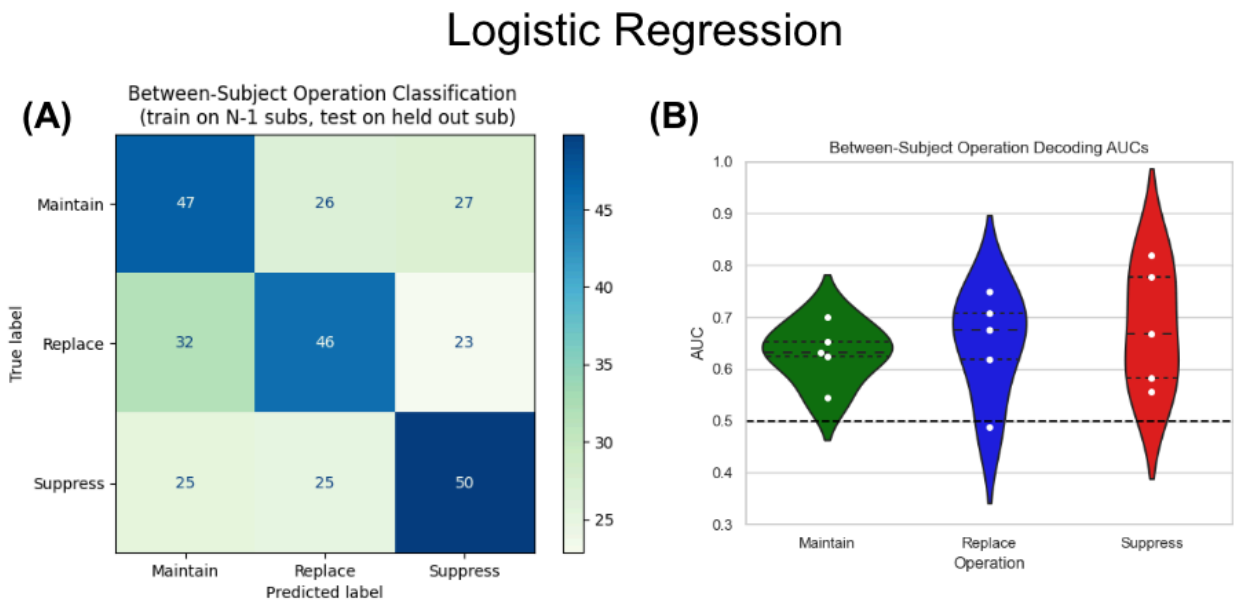

Figure 1. Logistic regression model. **(A)** Confusion matrix of between-subject classification chance = 33%). **(B)** ROC AUC values for between-subject classification (chance = 0.50).

Random Forest: Next, we implemented a Random Forest model with no pre-specified max depth. We found that the unconstrained model reached a max depth of 30, but considered that this may have led to overfitting due to the inherent variability of our data. Regardless, we found that each of the operations showed classification accuracy above chance (**Figure 2A**). Interestingly, we found that Maintain was the most accurately classified operation while Replace and Suppress, though classified above chance, were often inaccurately classified as Maintain. This may suggest that Replace and Suppress operations result in fundamentally more variable neural signals than maintain, leading to our RF model to inaccurately classify them when overfit on the noise of the training data. As demonstrated in **Figure 2B**, Replace and Suppress show much more variability and, though on average they can be classified above chance (AUC = 0.50), the bottom tails of each predict below-chance performance. Maintain, on the other hand, shows a much smaller range of AUC scores with nearly the entire distribution significantly above chance. The variability of the Replace And Suppress operations next led us to next test Random Forest models at several max-depth levels in order to avoid overfitting to better classify the three cognitive operations from each other.

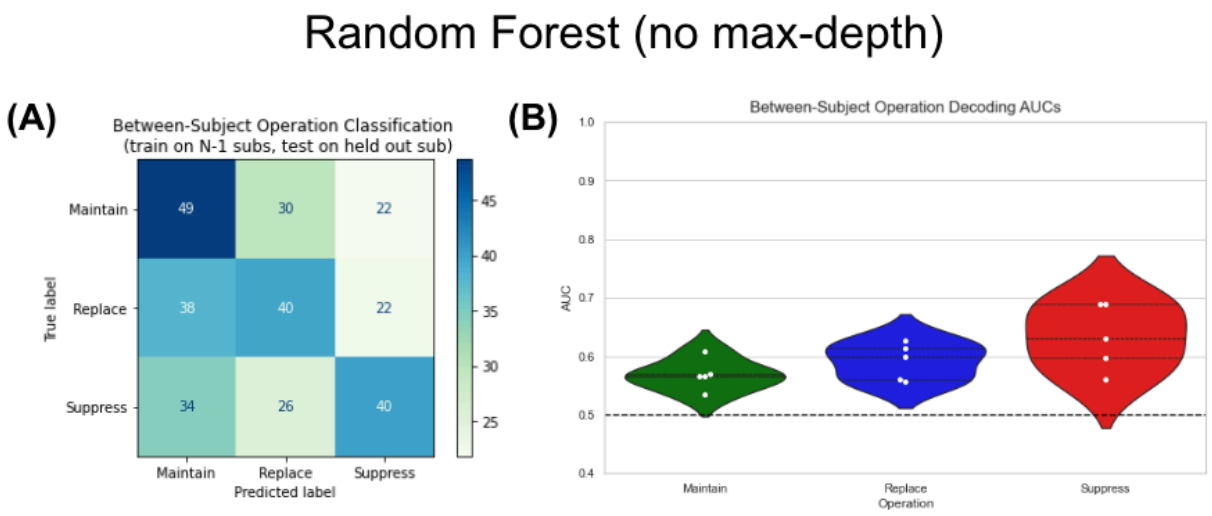

Figure 2. Random forest model (with no max-depth). **(A)** Confusion matrix of between-subject classification chance = 33%). **(B)** ROC AUC values for between-subject classification (chance = 0.50).

Constrained Random Forest: Since fMRI data is highly variable, we tested several Random Forest models with pre-specified max depths to avoid overfitting of our models. Due to computational constraints we did not iteratively test all possible max depth parameter values, but rather compared models of four categories: shallow, medium, deep, and unbounded. The shallow model was set to a max depth of 2 and resulted in mean AUC = 0.596. The medium model was set to a max depth of 10 and resulted in mean AUC = 0.608. The deep model was set to a max depth of 20, resulting in mean AUC = 0.602. The unbounded model, as described above, was not constrained in its decision points and resulted in a max depth of 30 with an average AUC of 0.607. While these mean AUC scores were quite similar suggesting that the number of decision points may not have a major effect on model performance, we selected the highest mean AUC model with max depth 10 to compare to other models.

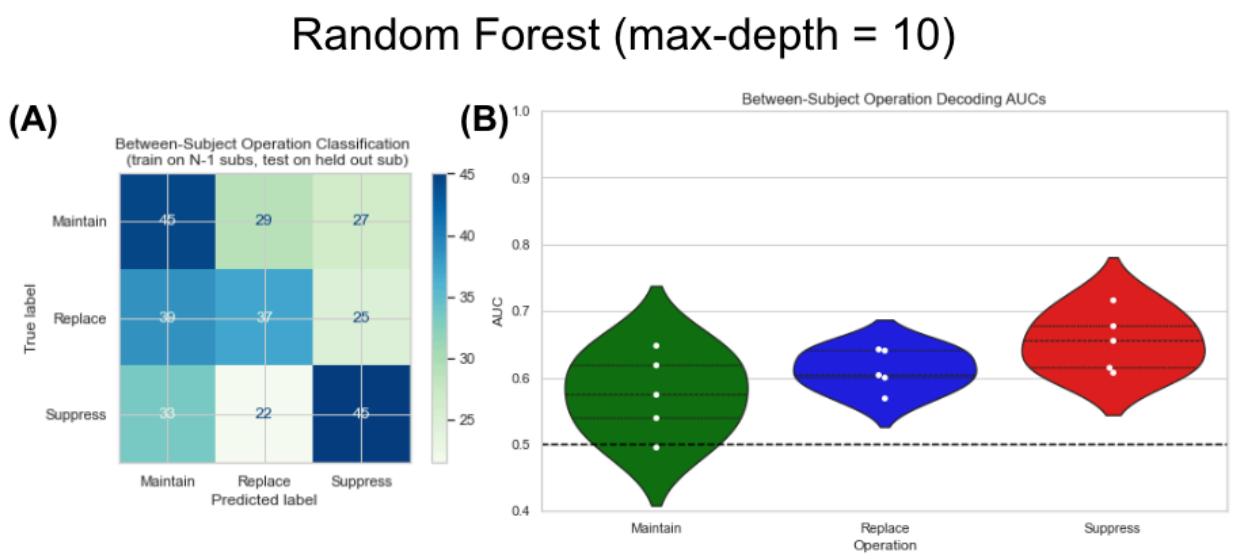

Figure 3. Random forest model (with max-depth of 10). **(A)** Confusion matrix of between-subject classification chance = 33%). **(B)** ROC AUC values for between-subject classification (chance = 0.50).

XGBoost: Building on the work based on the RF model, we trained an XGBoost model with a max depth of 10. Results of the confusion matrix are shown in **Figure 4A**, the hit rate is 53.4% for Maintain condition, 53.2% for Replace condition and 61.1% for Suppress condition. The model successfully classified each condition above chance (33%). Violin plots (**Figure 4B**) revealed all between-subject AUC all performed above chance (50%). Suppress condition has the highest AUC, compared to Replace and Maintain (estimated mean and 95% CI see below). There is also more variance in Suppress condition compared to Maintain and Replace.

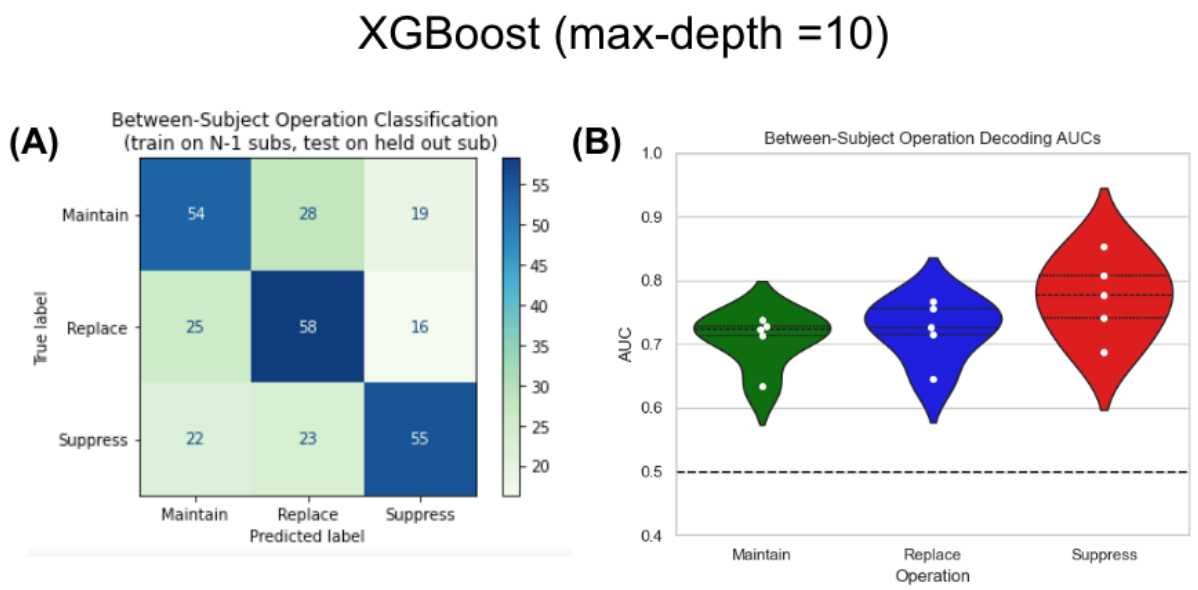

Figure 4. XGBoost model (with max-depth of 10). **(A)** Confusion matrix of between-subject classification chance = 33%). **(B)** ROC AUC values for between-subject classification (chance = 0.50).

Formal model Comparisons: We then asked if there is a task-specific effect on AUC score, such as one task having a higher AUC score than other tasks. Mixed ANOVA using condition as a between-subject variable and an error term of subject-variance, showed there was a near-significance main effect of condition ($F(1.14, 4.57) = 5.22$, $P = 0.07$), which suggests AUC score showed difference between some conditions. The estimated mean and 95% confidence interval for Maintain (M) is 0.708 [0.656, 0.760], for Replace (R) is 0.722 [0.663, 0.781], and for Suppress (S) is 0.773 [0.695, 0.852]. It is evident that the Suppress condition has a higher upper bound and more spread than the other two conditions. Pairwise comparison showed no difference between the conditions (M-R: $t(4) = -1.884$, $P = 0.257$, [-0.043, 0.0133]; M-S: $t(4) = -2.628$, $P = 0.12$, [-0.155, 0.0234], R-S: $t(4) = -1.952$, $P = 0.239$, [-0.144, 0.0421], Tukey method was employed for P value Type-I error adjustment).

Discussion

Logistic regression has been used in previous literature (e.g., Kim et al., 2020) to classify patterns of neural activity evoked during cognitive control tasks. Logistic regression is easy to implement and can perform well handling large datasets (such as neuroimaging data) with less computational demands compared to more complex models (e.g., Random Forest, XGBoost). We replicate previous work (Kim et al., 2020) in a novel set of data, demonstrating that distinct cognitive control strategies can be classified from distributed patterns of neural activity. While we were able to classify each operation above chance (**Figure 1A**), there is some confusion in the off-diagonal cells that approach chance level (e.g., Replace being confused as Maintain). This finding is expected and consistent with prior work (Kim et al., 2020). One limitation in the current project is the small sample size ($N = 5$). While each participant's data is very high-dimensional, this is less than a typical sample size for many modern fMRI studies. We would expect the main effects to remain consistent when increasing the sample size but the confusion between operations during classification, and variance in model performance, should be decreased with an adequately powered sample.

Our goal in implementing Random Forest models at several levels of depth was to consider the effect of model complexity on performance and at what level of depth we saw overfitting of our highly variable neuroimaging data. We saw comparatively similar but slightly poorer performance from our unconstrained Random Forest model (resulting in a max depth of 30), suggesting that there was a degree of overfitting when too many decision points were considered. Interestingly, the main consequence of this overfitting appeared to be more false positive results for the Maintain condition. This may be due to the relative decrease in variability for the Maintain condition compared to the Replace and particularly Suppress conditions. Perhaps the model struggled to distinguish Replace and Suppress because of an inherently

noisier neural signal while Maintain was less variable, resulting in improved performance as well as increased false positives for the Maintain operation.

While the unconstrained Random Forest model did perform above chance, our main goal in selecting the Random Forest approach was to test models at several levels of depth to determine the effect of model complexity on classification accuracy of the BOLD signal. After iteratively testing shallow, medium, and deep max depths, we found that our model performed best with a medium max depth of 10, suggesting that model performance was best when the model was allowed a degree of complexity, but limited from overfitting. Interestingly, the max depth 10 Random Forest model showed improved performance on classifying Suppression compared to the unconstrained Random Forest model. This suggests that the unconstrained model was indeed overfit on the variability of the Suppress data and that limiting decision points could avoid some of that overfitting. In addition, however, we saw greater variability in classification accuracy for the Maintain condition than we did with the unconstrained model. This may reflect that, because the Maintain signal was less variable, it benefitted from a more complex model with more decision points. The difference in classification accuracy between conditions based on the max depth of the model may suggest the inherent variability of the different cognitive processes and the difficulty or strategy differences we observe between subjects asked to perform complex cognitive operations. Ultimately the results of these two models reflect a tradeoff in the classification of complex, non-linear neuroimaging data where it is difficult to determine an optimal model that captures the complexity of the data while also avoiding overfitting. Notably, however, we did not iteratively test all possible max depth levels due to computational constraints, and a promising future direction would be to determine the optimal max depth based on a more comprehensive parameter optimization process.

The goal of the XGBoost model is to capture the variance in the BOLD data by a more complicated model structure. In our study, XGBoost was proved to be a better model than Random Forest model. Two reasons could contribute to such a result. Compared to the bagging method in the Random Forest model, the XGBoost model trains trees sequentially instead of in parallel, which gives an opportunity for correction when applying the current tree algorithm to new data. This is potentially better for noisy BOLD data to avoid overfitting. There is also evidence that the RF model is better for a bigger dataset while XGBoost can be applied for a smaller dataset, and our sample size is 5. Indeed, in another breast-cancer study with 275 instances, XGBoost and Random Forest models have similar accuracy (Kabiraj et al., 2020).

However, the accuracy of XGBoost is not high compared to other studies. A seminal study (Torlay et al., 2017) applied XGBoost to classify BOLD in five ROIs when patients are doing tasks with ~91% accuracy. One possible reason is that we employed classification based on whole-brain instead of restricted ROI, which could have higher signal-to-noise ratio. However, given we don't have specific hypotheses about which regions are more involved in these operations, applying classification to the whole-brain will reduce bias.

Despite better performance, both RF and XGBoost displayed a consistent pattern in between-subject decoding AUCs, which is the Suppress condition having more spread and higher AUC score than the other two conditions. This suggests Suppress condition is a more elaborate process and could result in more variability in strategies employed.

Conclusion

In conclusion, we show here that complex working memory operations of Maintaining, Replacing, and Suppressing visual information result in fundamentally different activity in the human brain. Testing four models, we demonstrate that these three operations can be

successfully classified at above chance accuracy. While all models successfully distinguished between the three operations, we further demonstrate that different classification approaches result in classification accuracy differences. Comparing an L2 logistic regression model, unconstrained random forest model, multiple constrained random forest models, and an XGBoost model, we show that XGBoost results in the most accurate classification while logistic regression and random forest models do not significantly differ. This suggests that XGBoost models may be optimal for classifying neuroimaging data based on their complexity and flexibility in handling non-linear, highly variable data. Furthermore, all models were most successful in classifying the Maintain operation and least accurate in classifying the Suppress operation, suggesting that suppression of observed information may be a more difficult or variable cognitive process than maintaining it in one's mind. Future work should continue to explore optimal classification approaches for neuroimaging data including comprehensive parameter optimization and cross-validation across a larger sample of subjects.

Acknowledgments

Group Member	**Contributions**	**Contribution Score**
Caleb Jerinic-Brodeur	● Acquired data, set-up data, cleaned data ○ Got all the data to the point that models could be implemented ● Implemented logistic regression model ● Ran formal model comparisons for all tested models ● Wrote Introduction and Data sections ● Contributed to writing of Modeling & Validation, Results, Discussion	100
Owen Friend	● Implemented random forest models ● Compared models by max depth ● Wrote/contributed writing to Hypotheses, Modeling/Validation, Results, Discussion, Conclusion	100
Lingwei	● Implemented XGBoost model	100

Ouyang	• Modeling/Validation, Results, Discussion	

References

Banich, M. T., Mackiewicz Seghete, K. L., Depue, B. E., & Burgess, G. C. (2015). Multiple modes of clearing one's mind of current thoughts: Overlapping and distinct neural systems. *Neuropsychologia*, *69*, 105–117. https://doi.org/10.1016/j.neuropsychologia.2015.01.039

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... & Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature methods*, *16*(1), 111-116. https://doi.org/10.1038/s41592-018-0235-4

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Kim, H., Smolker, H. R., Smith, L. L., Banich, M. T., & Lewis-Peacock, J. A. (2020). Changes to information in working memory depend on distinct removal operations. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-20085-4

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008

Lewis-Peacock, J. A. & Norman, K. A. Multivoxel pattern analysis of functional MRI data. In *The Cognitive Neurosciences* (eds Gazzaniga, M. S. & Mangun, G. R.). 911–920 (MIT Press, 2014).

Lewis-Peacock, J. A., Kessler, Y., & Oberauer, K. (2018). The removal of information from working memory. *Annals of the New York Academy of Sciences*, *1424*(1), 33–44. https://doi.org/10.1111/nyas.13714

Luck, S. J., & Vogel, E. K. (2013). Visual Working Memory Capacity: From Psychophysics and Neurobiology to Individual Differences. *Trends in Cognitive Sciences*, *17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain informatics*, *4*, 159-169.

Kabiraj, S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S. A., & Podder, E. (2020, July). Breast cancer risk prediction using XGBoost and random forest algorithm. In *2020 11th international conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-4). IEEE.