Application of Random Forest to classify EEG data of mTBI patients and control adults obtained during a Visuospatial Working Memory Task

Cruz, W.,[1] Cavanagh, J.F.,[2] Lin, C.Y.[1]

[1] [2]University of New Mexico, UNM

Introduction

The combination of electroencephalographic (EEG) recording and cognitive experimental tasks provides an excellent tool for studying human neural dynamics. EEG provides high temporal resolution time series data sampled across multiple scalp locations that produces large amounts of data, and some of these datasets are freely available in repositories on the internet (Fig. 1).

OpenNEURO BETA
A free and open platform for analyzing and sharing neuroimaging data

Figure 1. Open-science neuroinformatics database repository.

The analysis of EEG data requires the implementation of many signal extraction methods that are further used to characterize the psycho physiological phenomena been studied. Sometimes the large amount of data collected for one experiment is not fully used because most of the researchers prefer to take a confirmatory and deductive approach, thus focusing their attention in a rather narrow number of features but unintentionally overlooking other important latent features. The utility of the random forest algorithm was investigated for extracting the most relevant features from an EEG data set obtained from mild Traumatic Brain Injury (mTBI) patients and Healthy Controls (HC) during a Visuospatial Working Memory (VSWM) Task (Cavanagh, 2021).

Random forest is a supervised ML method used for classification purposes; the algorithm generates a set number of decision trees, each of which is made based on different subsets of data extracted from the training set. These subsets are selected following a random sampling approach; this iterative process and the number of decision trees computed are further used to reach a classification consensus, and the most common output is selected as the most relevant model which is usually the one that contains the most relevant features for the purpose of classifying the cases or instances considered (Liaw and Wiener, 2002). Therefore it is possible to classify EEG signal recording using this data analysis framework to characterize the neural dynamics and predict the performance and diagnosis (Klimesch, 1997).

Advantages of Analyzing EEG data with Random Forest

Machine Learning techniques include different computational frameworks such as SVM, Logistic Regression and Random Forest, that are capable of mining large datasets, some advantages of using these frameworks and in particular Random Forest when analyzing EEG datasets are:

- **Identify** relevant questions concerning EEG data.
- **Discover** new knowledge through pattern recognition and mathematical modeling.
- **Applicability** in both medical and social sciences' datasets.
- **Predict** the performance in a task.
- **Diagnose** patients based on their neural activity patterns.

Methods

Subjects

An EEG data set obtained during a Visuospatial Working Memory (VSWM) task was downloaded from OpenNeuro (Cavanagh, 2021); the subjects that conformed the final groups for the present analysis, mild Traumatic Brain Injury patients (mTBI, n = 27) and Healthy Controls (HC, n =27), were matched using demographic variables and their scores in the task, thus ensuring that both groups did not differ significantly by age (p = 0.67), sex (p = 0.58), nor by hit ratio (p= 0.97).

Procedure

The VSWM task (see Figure 2) was deployed using MATLAB, participants had to perform a yes-no recognition task and were asked to respond whether a location defined by a square containing a question mark (i.e. probe) had been occupied by a red dot in the preceding visual array. Participants were told to ignore the yellow dots. Altogether there were three conditions, either showing three targets or three red dots (*Condition 3*), showing three targets and two distractors (*Condition 3 + 2*), or showing 5 targets (*Condition 5*).

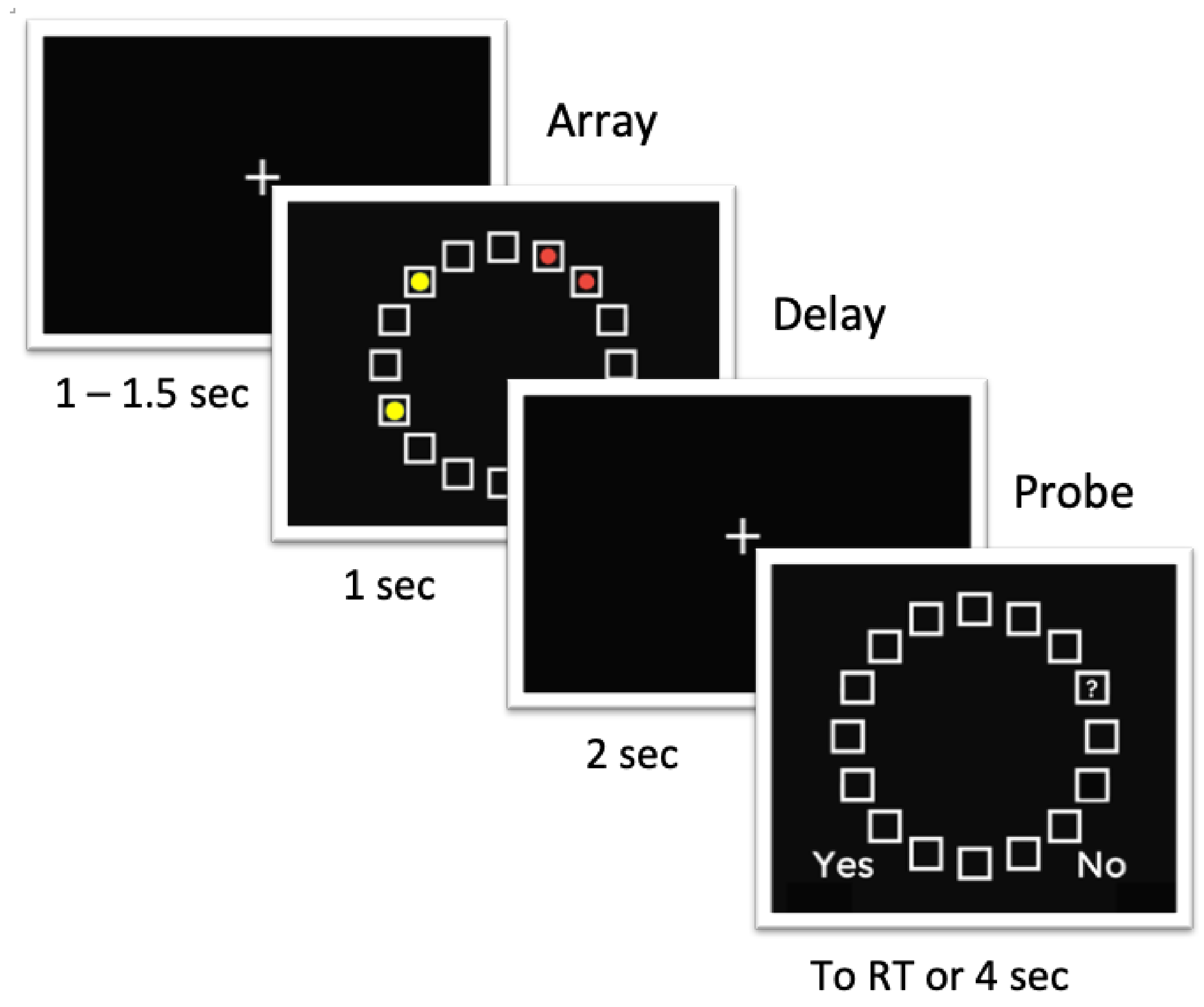

Figure 2. Visuospatial Working Memory (VSWM) task.

EEG Data

EEG activity was recording using a 64 channel cap and the electrodes were located according to the 10-20 system. Continuous EEG was monitored and reference to the *Fz* channel and acquired at a 1000 Hz sampling rate. Total EEG set up time was in average 30 minutes, and the VSWM task included a practice block and an experimental block with 150 trials. The data of each subject that comprised the final sample was first loaded into MATLAB using EEGLab (Delorme and Makeig, 2004), then the EKG and VEOG were removed leaving only the channels that recorded brain activity.

Following, each subject's data was decomposed by ICA using the picard algorithm (Ablin et al., 2018) and the artifact components were removed with the SASICA module (Chaumon et al., 2015). Lastly, the data was epoched including a period from -4000 to 0 ms (i.e. 0 was the time in which the probe appeared on the screen), thus covering three memory phases (i.e. Baseline, Encoding, Retention). This data was analyzed to extract 5 frequency components from each of the scalp sites. The EEG data was labelled by group and separated as either correct or incorrect for classification purposes independently of the condition, and was exported into plain format for further processing with R (R Core Team, 2020). The Random Forest Algorithm was trained with 60% of the data to build EEG classifiers of VSWM trial accuracy and diagnosis.

References

Ablin, P., Cardoso, J.-F., and Gramfort, A. (2018). Faster independent component analysis by preconditioning with hessian approximations. *IEEE Transactions on Signal Processing*, 66(15):4040–4049.

Baddeley, A. D. and Hitch, G. (1974). Working memory. In *Psychology of learning and motivation*, volume 8, pages 47–89. Elsevier.

Boonstra, T. W., Powell, T. Y., Mehrkanoon, S., and Breakspear, M. (2013). Effects of mnemonic load on cortical activity during visual working memory: linking ongoing brain activity with evoked responses. *International journal of psychophysiology*, 89(3):409–418.

Brandt, E., Wilson, J. K., Rieger, R. E., Gill, D., Mayer, A. R., and Cavanagh, J. F. (2020). Respiratory sinus arrhythmia correlates with depressive symptoms following mild traumatic brain injury. *Journal of Psychophysiology*.

Broadway, J. M., Frank, M. J., and Cavanagh, J. F. (2018). Dopamine d2 agonist affects visuospatial working memory distractor interference depending on individual differences in baseline working memory span. *Cognitive, Affective, & Behavioral Neuroscience*, 18(3):509–520.

Broadway, J. M., Rieger, R. E., Campbell, R. A., Quinn, D. K., Mayer, A. R., Yeo, R. A., Wilson, J. K., Gill, D., Fratzke, V., and Cavanagh, J. F. (2019). Executive function predictors of delayed memory deficits after mild traumatic brain injury. *cortex*, 120:240–248.

Cavanagh, J. F. (2021). ”eeg: Visual working memory in acute tbi”.

Cavanagh, J. F., Rieger, R. E., Wilson, J. K., Gill, D., Fullerton, L., Brandt, E., and Mayer, A. R. (2020). Joint analysis of frontal theta synchrony and white matter following mild traumatic brain injury. *Brain imaging and behavior*, 14(6):2210–2223.

Chaumon, M., Bishop, D. V., and Busch, N. A. (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of neuroscience methods*, 250:47–63.

Daróczi, G. and Tsegelskyi, R. (2021). *pander: An R 'Pandoc' Writer*. R package version 0.6.4.

Delorme, A. and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21.

Firke, S. (2021). *janitor: Simple Tools for Examining and Cleaning Dirty Data*. R package version 2.1.0.

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition.

Fox, J., Weisberg, S., and Price, B. (2020). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-4.

Helwig, N. E. (2014). *eegkitdata: Data for package eegkit*. R package version 1.0.

Helwig, N. E. (2018a). *bigsplines: Smoothing Splines for Large Samples*. R package version 1.1-1.

Helwig, N. E. (2018b). *eegkit: Toolkit for Electroencephalography Data*. R package version 1.0-4.

Helwig, N. E. (2018c). *ica: Independent Component Analysis*. R package version 1.0-2.

Hindriks, R. and van Putten, M. J. (2013). Thalamo-cortical mechanisms underlying changes in amplitude and frequency of human alpha oscillations. *Neuroimage*, 70:150–163.

Howard, M. W., Rizzuto, D. S., Caplan, J. B., Madsen, J. R., Lisman, J., Aschenbrenner-Scheibe, R., Schulze-Bonhage, A., and Kahana, M. J. (2003). Gamma oscillations correlate with working memory load in humans. *Cerebral cortex*, 13(12):1369–1374.

J, L. (2006). Plotrix: a package in the red light district of r. *R-News*, 6(4):8–12.

Johannesen, J. K., Bi, J., Jiang, R., Kenney, J. G., and Chen, C.-M. A. (2016). Machine learning identification of eeg features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatric electrophysiology*, 2(1):1–21.

Kane, M. J. and Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review*, 9(4):637–671.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.4.0.

Kelley, D. and Richards, C. (2021). *oce: Analysis of Oceanographic Data*. R package version 1.4-0.

Kelley, D., Richards, C., and SCOR/IAPSO, W. (2017). *gsw: Gibbs Sea Water Functions*. R package version 1.0-5.

Klimesch, W. (1997). Eeg-alpha rhythms and memory processes. *International Journal of psychophysiology*, 26(1-3):319–340.

Klimesch, W. (1999). Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2-3):169–195.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-88.

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.4-0.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2021). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-7.

Murdoch, D. and Adler, D. (2021). *rgl: 3D Visualization Using OpenGL*. R package version 0.106.8.

Müller, K. and Wickham, H. (2021). *tibble: Simple Data Frames*. R package version 3.1.2.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rinker, T. W. and Kurkiewicz, D. (2018). *pacman: Package Management for R*. Buffalo, New York. version 0.5.0.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.

Schauberger, P. and Walker, A. (2020). *openxlsx: Read, Write and Edit xlsx Files*. R package version 4.2.3.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Crowley, J. (2021). *GGally: Extension to 'ggplot2'*. R package version 2.1.1.

signal developers (2014). *signal: Signal processing*.

Stanley, D. (2021). *apaTables: Create American Psychological Association (APA) Style Tables*. R package version 2.0.8.

Vis, J. (2019). *edfReader: Reading EDF(+) and BDF(+) Files*. R package version 1.2.1.

Wickham, H. (2011a). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Wickham, H. (2011b). testthat: Get started with testing. *The R Journal*, 3:5–10.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.

Wickham, H. (2021a). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.5.1.

https://github.com/billytaipei101 Vision Sciences Society Symposia 2022, Florida FL wi.cruzm@gmail.com