Research Topic: Predicting Deficits in Action Prediction from Tumor Location

Requirement: Fall 2025, Stat 627, Statistical Machine Learning

Team: Sally Henley, Ogechi Onyewu, Rebecca Tegiacchi

Questions of Interest:

1. Can we predict performance accuracy on an action prediction task based on lesion location within the brain (infratentorial vs supratentorial)?
2. Does age at diagnosis predict performance on the action prediction task or general cognitive performance?

Planned Approach:

The cerebellum is a key developmental structure that has been shown to be highly involved in both motor and cognitive functioning. Despite only being 10% of the volume of the brain, the cerebellum contains over half of the brain's neurons.

Literature Review:

Additionally, the cerebellum has been divided into functional subregions, such that areas associated with motor coordination are located medially, in the vermis, while areas that support cognitive functioning are typically located laterally (Stoodley et al., 2016). Importantly, the cerebellum has been shown to be functionally connected to key cortical areas, such as prefrontal, temporal, and parietal association areas, which are known for functions such as planning, decision making, and language, among others (Stoodley & Schmahmann, 2010). Disruption to cerebellar subregions can be detrimental to cognitive and motor functioning, whether due to tumor, stroke, or atrophy. The universal cerebellar transform theory has been proposed, stating that the cerebellum uses prediction and error signaling from inputs using the cerebro-cerebellar loops, integrating this information and generating internal models that are used to predict actions in multiple domains, including language, affect, and action processing (Leggio & Molinari, 2015; Argyropoulos, 2016).

The cerebellum is extremely important in development. Cerebellar volume peaks between 12-16 years old, which is much later than the timeline of peak cerebrum development (Tiemeier et al., 2010). Unlike the rest of the brain, where early damage typically leads to better outcomes, it is hypothesized that damage to the cerebellum at an early age actually leads to worse long-term outcomes (Scott et al., 2001; Starowicz-Filip et al., 2017). The cerebro-cerebellar circuits are necessary for learning and skill development, as well as encoding the internal models that are necessary for the control of movement and mental representations (Ito, 2008; Stoodley & Limperopoulos, 2016).

Machine learning is not a new method in neuroscience research. Glaser et al. (2019) discusses the various uses of machine learning, highlighting the prediction of continuous and categorical output using regression and classification. Because the variables used in clinical research are usually not completely separate and distinct, machine learning methods determine if variables can predict each other, especially when the variables aren't necessarily linear. The various regression methods that will be used in this project have been used extensively in clinical neuroscience research (see Xie et al., 2016; Hoffman et al., 2015; Ritsner & Strous, 2010 for example).

Data Assessment:

The data for the current project is from a study conducted by Butti et al. (2020), in which researchers were analyzing the impact of lesion location on predictive accuracy, specifically taking age at diagnosis into account. In this study, participants were shown videos that showed a child performing a task in which they were reaching to grasp an object. Then, they were presented with shortened versions of the same videos and had to predict or infer the outcome. They had results from N = 63 patients. To analyze the impact of lesion location, they split patients into control, supratentorial lesions, and infratentorial lesions, with N = 21 patients in each group. The supratentorial lesions are defined as being *outside of the cerebellum*, typically in the cortex, while infratentorial lesions are *within the cerebellum*. For the current analysis, we will use Accuracy Testing as the primary response variable, with predictors including lesion location, age at diagnosis, time since diagnosis, tumor type, FISQ (measure of general cognitive ability), and Social Perception scores. We will treat radiation, chemotherapy, and neurosurgery as covariates. We will also look at the relationship between lesion location on

the beta index, which is a measure of the strength of the contextual priors (the degree to which expectations from prior observations of a specific event predict future events).

Dataset Overview:

Dataset Source: Cerebellar Damage Affects Contextual Priors for Action Prediction in Patients with Childhood Brain Tumor

Number of Observations = 63 total

- 42 combined observations:
  - 21 for Supratentorial and
  - 21 for Infratentorial
- 21 for Control Group

Response Variables:

1) Beta Index: Measurement of Response Bias (how likely to make the same association per the probability of occurrence)
2) Accuracy Testing phase percentages: Associations between contextual clues and actions (preestablished probability of co-occurrence)

Predictors:

1) Tumor type
2) Tumor location (supratentorial vs infratentorial)
3) Age at diagnosis (months)
4) Time since diagnosis (months)
5) Presence of radiotherapy, chemotherapy, or neurosurgery (covariates)
6) FSIQ (measure of general cognitive capacity)
7) Social Perception Scores

Planned SML methods:

For Data cleaning and formatting, we remove empty rows, merge data, and format as csv. We normalize features as scaling is important and will be needed for KNN analysis. The project team identified Statistical Machine Learning methods which will be utilized and leveraged to gain data-driven insight to answer proposed questions. Below are the proposed methods and logic overview based on IDA/EDA of the data set.

Logistic/Linear Regression: Generate generalize model; We already suspect collinearity between clinical information for presence of radiotherapy, chemotherapy, and neurosurgery and tumor type and to address this, we will analyze the VIF and generate covariance matrices and heatmaps to detect and determine which predictors are most correlated.

KNN: We will leverage KNN to explore non-linear relationships between features and target variables, which logistic regression may not capture effectively. Compare KNN with logistic regression and PLS for interpretability. Compare MSE with tuned Ridge/LASSO MSE.

Jackknife: We will use this resampling method to estimate stability and variance/bias for the KNN model and for further model evaluation/validation.

Principal Component Analysis [PCA]: Leverage for dimensionality reduction; transform the initial set of predictors with identified multicollinearity to a smaller set of uncorrelated variables called principal components. These principal components will feed the PCR method. We will test implementing PCA before KNN if we have high collinearity between predictors to reduce correlated features and noise.

K-Fold Cross Validation: We will perform this method to analyze cross-validation loops for stacking (stacked model that combines KNN and Ridge/LASSO components) and collect MSE on each fold. We will test k = 5 and k = 10.

Ridge/LASSO: We will use Ridge to handle multicollinearity and LASSO to perform feature selection. Use PCA reduced features and KNN predictions as input.

Principal Component Regression [PCR]: We will use this method to find new features or add new variables.

Partial Least Squares [PLS]: We will use this method after PCR, since PLS considers the response variable(s) during component extraction and maximizes covariance between predictors and response variable(s), which may lead to a better predictive performance than PCR.

Ethical Concerns and Risks:

Though this data uses medical data, all patient information has been removed, and there are no instances of personally identifiable health information. There are no risks associated with analyzing the data.

Deliverable:

The team will present a poster summarizing research analysis and outcomes. Detailed below is an overview of the poster format.

- Background on research topic
- Neuroscience Term Definition
- Data Set Description
- Data Visualizations and summary of observations
- Summary of Results/Conclusions
- Recommendations

Schedule and Hours:

Week	Main Activities	Team Estimated Hours	Expected Outcomes
Week 10: Oct. 26 – Nov 1, 2025	Finish cleaning/organizing data, start modeling	12	- Have data set up and ready for analysis - Start modeling (linear and logistic regression, Jackknife)
Week 11: Nov 2 – 8, 2025	Finish data modeling	12	- Finish rest of modeling (KNN, PCA/PCR, PLS)
Week 12: Nov 9 – 15, 2025	Start poster, finalize models	12	- Finish background/methods of paper - Finish all modeling
Week 13: Nov 16 – 22, 2025	Interpret results, work on poster	12	- Make poster figures, organize results and make progress on poster
Week 14: Nov 23 – 29, 2025	Finish poster, prepare for presentation, put items on GitHub	12	- Finish poster - Prepare for presentation (dry run practices)

Group Member Responsibilities:

Rebecca: Research, review, and analyze various published articles dealing with neural activity and neuroimaging data to discuss with team to finalize data set. Clean and organize dataset to prepare for analysis. Utilize neuroscience acumen and expertise to help the team identify final data set. Draft initial Project Plan, literature review, and data assessment and walkthrough with team for understanding and to refine and finalize Project Plan. Identify, source, review/analyze appropriate references for research. Analyze data and implement linear and logistic regression to generate models and visualizations for further analysis. Set up formatting for poster, draft and edit background and results.

Ogechi: Research, review, and analyze various published articles dealing with neural activity and neuroimaging data to discuss with team to finalize data set. Review and read reference documents to gain better insight into research field and to gain better understanding of neuro terminology. Analyze linear/logistic models and KNN to implement Jackknife resampling. Ridge/LASSO, and PCA/PCR/PLS methods. Update/refine Planned SML Methods section of Project Plan. Migrate research artifacts to GITHUB site. Draft and edit analysis plan and results on poster. Organize, store, and track all Research Project Team artifacts in one location and use this to discern upload to future GITHUB site.

Sally: Review team findings and corroborate with team to finalize data set. Continue to clean and format data per conducted IDA and EDA. Document initial questions and review with team to gain more insight and direction for generating models and communicating initial and exploratory analysis with the team. Analyze linear and logistic regression models and implement KNN classification to generate models and visualizations for further analysis. Review minimum of two identified references from list to gain better insight into research field and to gain better understanding of neuro terminology. Collaborate team for KNN preprocessing. Draft and edit analysis plan and results on poster.

All: Participate in weekly recurring meetings to edit and finalize project plan, propose and formulate project plan timeline, discuss and analyze findings, draft, edit, and finalize poster, propose and formulate project plan timeline.

Summary:

We will use data collected by Butti et al. (2020) to examine the differences in performance on action prediction tasks based on lesion location within the brain. We will use methods such as linear and logistic regression, Ridge/LASSO regression, KNN, jackknife, and PCR and PLS while using cross validation techniques to assess flexibility and find the best set of predictors. We will compile our models to come up with a final model that creates the best representation of the data. We will use the results from our models to assess the impact of lesion location on task performance and discuss future research that could be conducted to continue to make advances in the field.

References

Argyropoulos G. P. (2016). The cerebellum, internal models and prediction in 'non-motor' aspects of language: A critical review. *Brain and language*, *161*, 4–17. https://doi.org/10.1016/j.bandl.2015.08.003

Butti, N., Corti, C., Finisguerra, A., Bardoni, A., Borgatti, R., Poggi, G., & Urgesi, C. (2020). Cerebellar damage affects contextual priors for action prediction in patients with childhood brain tumor. *Cerebellum (London, England)*, *19*(6), 799–811. https://doi.org/10.1007/s12311-020-01168-w

Glaser, J. I., Benjamin, A.S., Farhoodi, R., Kording, K.P. (2019). The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology, 175,* 126-137. https://doi.org/10.1016/j.pneurobio.2019.01.008.

Hoffman, H., Lee, S. I., Garst, J. H., Lu, D. S., Li, C. H., Nagasawa, D. T., Ghalehsari, N., Jahanforouz, N., Razaghy, M., Espinal, M., Ghavamrezaii, A., Paak, B. H., Wu, I., Sarrafzadeh, M., & Lu, D. C. (2015). Use of multivariate linear regression and support vector regression to predict functional outcome after surgery for cervical spondylotic myelopathy. *Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia*, *22*(9), 1444–1449. https://doi.org/10.1016/j.jocn.2015.04.002

Ito M. (2008). Control of mental activities by internal models in the cerebellum. *Nature reviews. Neuroscience*, *9*(4), 304–313. https://doi.org/10.1038/nrn2332

Leggio, M., & Molinari, M. (2015). Cerebellar sequencing: a trick for predicting the future. *Cerebellum (London, England)*, *14*(1), 35–38. https://doi.org/10.1007/s12311-014-0616-x

Ritsner, M. S., & Strous, R. D. (2010). Neurocognitive deficits in schizophrenia are associated with alterations in blood levels of neurosteroids: a multiple regression analysis of findings from a double-blind, randomized, placebo-controlled, crossover trial with DHEA. *Journal of psychiatric research*, *44*(2), 75–80. https://doi.org/10.1016/j.jpsychires.2009.07.002

Scott, R. B., Stoodley, C. J., Anslow, P., Paul, C., Stein, J. F., Sugden, E. M., & Mitchell, C. D. (2001). Lateralized cognitive deficits in children following cerebellar lesions. *Developmental medicine and child neurology*, *43*(10), 685–691. https://doi.org/10.1017/s0012162201001232

Starowicz-Filip, A., Chrobak, A. A., Milczarek, O., & Kwiatkowski, S. (2017). The visuospatial functions in children after cerebellar low-grade astrocytoma surgery: A contribution to the pediatric neuropsychology of the cerebellum. *Journal of neuropsychology*, *11*(2), 201–221. https://doi.org/10.1111/jnp.12093

Stoodley, C. J., & Limperopoulos, C. (2016). Structure-function relationships in the developing cerebellum: Evidence from early-life cerebellar injury and neurodevelopmental disorders. *Seminars in fetal & neonatal medicine*, *21*(5), 356–364. https://doi.org/10.1016/j.siny.2016.04.010

Stoodley, C. J., MacMore, J. P., Makris, N., Sherman, J. C., & Schmahmann, J. D. (2016). Location of lesion determines motor vs. cognitive consequences in patients with cerebellar stroke. *NeuroImage. Clinical*, *12*, 765–775. https://doi.org/10.1016/j.nicl.2016.10.013

Stoodley, C. J., & Schmahmann, J. D. (2010). Evidence for topographic organization in the cerebellum of motor control versus cognitive and affective processing. *Cortex; a journal devoted to the study of the nervous system and behavior*, *46*(7), 831–844. https://doi.org/10.1016/j.cortex.2009.11.008

Tiemeier, H., Lenroot, R. K., Greenstein, D. K., Tran, L., Pierson, R., & Giedd, J. N. (2010). Cerebellum development during childhood and adolescence: a longitudinal morphometric MRI study. *NeuroImage*, *49*(1), 63–70. https://doi.org/10.1016/j.neuroimage.2009.08.016

Xie, Q., Wang, S., Zhu, J., Zhang, X. (2016). Modeling and predicting AD progression by regression analysis of sequential clinical data. *Neurocomputing, 195*. 50-55. https://doi.org/10.1016/j.neucom.2015.07.145.