**Two ideas for framing:**

1. Searching for optimal model to most powerfully recognize, if present, the relationship between early-life stress and brain development
   1. In doing so, we reveal that the effect is robust to several modeling approaches.
2. In building probabilistic models, however, we do not interpret the pattern of differences between groups in a way meaningful for behavior
   1. Differences between groups may reflect different developmental patterns that are each in their own right adaptive.
      1. Cite several sources suggesting that increase and decrease in volume may differ by brain region, and it is still unknown what function increases and decreases in volume serve for behavior
   2. Probabilities also involve a degree of uncertainty, and so these results highlight that follow early-life stress, while powerful pattern recognition models can detect idiosyntractic brain development, growth outcomes for individuals are not guaranteed and are likely impacted by many factors.

**NOTE:** possibly move second row in plots to supplemental. Make y-axis more interpreted language (not just like, this is the quantity on the axis).

**Introduction** (outline)

* Efforts to replicate explanatory findings in psychology and neuroscience have lead to mixed results, through which some widely accepted theories have come into question
* One issue is that the most common statistical methods in the fields have been unsuitable for many of the studies in which they are applied
  + Explanatory models are widely used to generate predictive conclusions
  + Predominantly linear parametric models miss non-linear and higher-order effects
  + The high-dimensionality that comes with fMRI can produce biased regression estimates
  + (rebuttal) Though, simple parametric models are effective in the context of small datasets that has been most often the norm in due to resources
* It is worth evaluating if data-driven models replicate the conclusions of past work
* At the same time, data-driven approaches and predictive models are not without their own shortcomings
  + prone to overfitting
  + biased in the context of smaller samples with class imbalance
  + often require a lot of data to fit an algorithm
* Recently, it has been proposed that the field could benefit from taking a ‘converging evidence’ approach: researchers look across and within-studies for converging statistical evidence on which to base conclusions (Rocca & Yarkoni, 2020)
* By analyzing the consistency of results from models requiring varying “researcher degrees of freedom”, we can demonstrate the robustness of our findings to the sources of bias that usually go unreported
* This paper takes such an approach to examine the effect of early-life adversity on subcortical grey matter development by conducting hypothesis-driven and hypothesis-free (data-driven) statistical analyses
* (background on the existing theories of early-life adversity’s impact on brain development)
  + Highlight that most of this work would fall into the theory-driven bin
* NOTE: use language in terms of probability
* No attempts to decode the impact of early-life adversity on brain development using data-driven machine learning exist, nor have they been rigorously compared within the same exact population and sample to the conclusions derived from traditional statistical methods

**Results**

*Model Accuracy and the reliability of associations between early-life adversity and subcortical brain volume.*

We built 6 ensemble models to predict group labels (probability of having experienced early-life adversity) from subcortical grey matter volume: 2 model types (gradient boosted machines (GBM) and logistic regression (GLM)) and 3 preprocessing pipelines (no covariate adjustment, adjustment for age and sex, and adjustment for age, sex, and cranial volume) for a total of 100 models of each kind and 600 models overall. Final results are derived by averaging the predicted values for each subject in the validation population, and calculating model performance one time for each pipeline x model-type (6 final prediction accuracies).

Diagram

Description automatically generated5 of the 6 models predicting group labels from brain volume were significantly more accurate than chance (highest ROC-AUC is .76, lowest **significant** is .66), while the only non-significant model is the GBM model using the age, sex, and cranial volume adjustment preprocessing pipeline, ROC-AUC: .61 (n.s.).

**Figure 1**. The plot contains 12 point clouds/boxplots (positioned in 6 cells), in which points are ROC-AUC values plotted on a y-axis ranging from around .35 to .85. The 6 point clouds on the top row (in the top 3 cells) correspond to the 6 final ROCAUC derived from averaging the predicted probabilities for each subject within each ensemble, and then calculating ROC-AUC one time (for each ensemble). The encoloured point clouds are derived from bootstrap resampling the 100 models of each ensemble, recomputing the averaged predicted probabilities and corresponding ROCAUC, 1000 times. The grey point cloud is the permuted null distribution used to compute the p-value for the 6 observed values. For a deeper analysis of the certainty and invariance across the 100 iterations of cross-validation separately, we calculate ROC-AUC *before* averaging predictions (100 distinct times) for each of the 6 ensembles, and plot the resulting 100-point cloud and boxplot in the bottom row (bottom 3 cells) for each of the 6 model-type by preprocessing-type combinations. Each column (and color) pertains to a different preprocessing pipeline, while within each cell, the cloud/boxplot on the left corresponds to the GBM models and ensemble and those on the right pertain to the GLM models and ensemble. Red lines reflect 95% confidence intervals, based either on bootstrap standard error (top row, encoloured point clouds), or span the empirical 5th to 95th quantiles (for the null and individual-fold distributions).

*Distinguishability of the Groups*

Quantifying the (dis)similarity between the ensemble’s predictions for participants of the two groups in the validation set can be a better way to evaluate the performance of the models than classification accuracy (or metrics based upon it such as ROC-AUC), because the thresholds (usually .5) required to make classifications are arbitrary unless the group sizes themselves carry some meaning (e.g., reflect the prevalence of a disease in a specific population). Thus, to assess the distinctiveness of each group’s subcortical brain features (with respect to the other’s), we also computed the Kolmogorov-Smirnov (KS) test statistic between the distribution of predictions made for the early-adversity and comparison groups within each ensemble. The Kolmogorov-Smirnov test is parametric, and enables statistical significance testing. However, to compliment our comparison of consistency between data- and theory-driven analyses, we also computed the empirical null distributions for the KS statistics calculated with each ensemble and a computed the permutation test p-value as well.

Figure 2 shows plots the empirical cumulative distribution function for each ensemble’s predictions for each of the two groups. Figure 3 plots the KS test statistic derived from comparing the two distributions (between participant groups) within each ensemble. Analogously to the plot of ROCAUC, the three plots in the top row of the figure show the KS statistic from the predicted values averaged across the 100-model ensemble while the bottom three plots show KS computed for each constituent model’s predictions. The results show that 4 of the 6 ensembles generated significantly different predictions for the two groups: both the GBM and GLM models using the unadjusted and age-and-sex-adjusted brain volumes. This also implies that after adjusting for cranial volume, neither ensemble successfully distinguished the two groups more significantly than would be expected by chance. The greatest vertical distance between the two distributions is generated by the GBM ensemble using unadjusted predictors (KS-statistic = .534, parametric p-value = .0002, permutation p-value < .001) followed by the GBM model using predictors adjusted for age and sex (KS = .472, parametric p-value = .0008, permutation p-value < .001), the GLM ensemble with unadjusted features (KS = .351, parametric p-value = .04, permutation p-value = .03) and finally the GLM model using predictors adjusted for age and sex (KS = .341, parametric p-value = .03, permutation p-value = .03).

***Chart, line chart, scatter chart

Description automatically generated***

\*

\*

\*\*\*

\*\*\*

**Figure 2**. The empirical cumulative distribution of predicted probabilities for each of the six ensembles, for each of the groups. Each column corresponds to a preprocessing pipeline, in which the first two rows correspond to the GBM using that pipeline, and the second two the GLM. Each dot represents a subject, and their position along the x-axis is the predicted probability (of belonging to the early-adversity group) assigned by each ensemble. The length of the bar subsuming their point is the 95% CI interval based on the variability in their predicted probabilities generated by the 100 models in each ensemble. Bars are colored based on the predicted probability of the ensemble and correspond to the x-axis of each plot, where red indicates lower probability, blue is higher, and green is the midpoint. The y-axis is the cumulative density for the distribution of probabilities assigned by each model to each group. As such, the y-axis (also) ranges from 0 to 1, and reflects the quantiles associated with each predicted probability plotted on the x-axis, for each group.

Chart, box and whisker chart

Description automatically generated

**Figure 3**. The plot contains 12 point clouds/boxplots (positioned in 6 cells), in which points are Kolmogorov-Smirnov test statistics. Values plotted on a y-axis ranging from around .2 to .65. The 6 point clouds on the top row (in the top 3 cells) correspond to the 6 final KS statistics derived from averaging the predicted probabilities for each subject from the 100 models in each ensemble, and then calculating the KS statistic one time for each. The encoloured point clouds are derived from bootstrap resampling the 100 models of each ensemble, recomputing the averaged predicted probabilities and corresponding KS statistic, 1000 times. The grey point cloud is the permuted null distribution used to compute the p-value for the 6 observed values. For a deeper analysis of the certainty and invariance across the 100 iterations of cross-validation separately, we calculate the KS statistic *before* averaging predictions (100 distinct times) for each of the 6 ensembles, and plot the resulting 100-point cloud and boxplot in the bottom row (bottom 3 cells) for each of the 6 model-type by preprocessing-type combinations. Each column (and color) pertains to a different preprocessing pipeline, while within each cell, the cloud/boxplot on the left corresponds to the GBM models and ensemble and those on the right pertain to the GLM models and ensemble. Red lines reflect 95% confidence intervals, based either on bootstrap standard error (top row, encoloured point clouds), or span the empirical 5th to 95th quantiles (for the null and individual-fold distributions).

*Discovery of Meaningful Subcortical Predictors via Ensemble Importance*

To evaluate which subcortical volumes were most important and generalizable for decoding the impact of early-adversity on subcortical brain volume, we computed the quantified the contribution that each variable made to the performance of each ensemble (ensemble importance; EI). The EI of each variable is the percent by which the ensemble ROCAUC (computed with the test data) is affected by replacing a variable with noise. The variable is replaced with noise (in the test data) and the predictions of each model in the entire ensemble are recomputed, the averaged vector of predicted probabilities are recalculated, and the ensemble ROCAUC is again assessed. This is repeated for each variable, and for each ensemble. To assess the significance of such values, the process is repeated 1000 times for each variable in each ensemble after first also randomly permuting the group labels in the validation population.

*Chart, scatter chart

Description automatically generated*As shown in Figure 4, In all 5 of the 6 ensembles that generated statistically significant ROCAUC, the left hippocampus emerged as the strongest predictor of group labels, and the left caudate the second-most predictive variable with the largest and second-largest EI, respectively. In permutation tests, these two variables are significant at an alpha level of .05 in all such models. Together this demonstrates the robustness of these predictors in hypothesis- and data-driven contexts, in which statistical significance is inferred from accurate model generalization. Other variables that reliably accumulated EI included the right ventral diencephalon, right accumbens, left pallidum and right caudate.

**Figure 4**. The plot contains one cell for each of the 6 ensembles (model-type by preprocessing-type combinations). Color-coding corresponds to the preprocessing pipeline, while the top row of plots pertain to the GBM ensembles, and the GLM ensembles for the same preprocessing pipeline plotted directly below. The x-axis plots the percent, positively or negatively, by which the ROCAUC of the ensemble computed in the validation population changes after randomly permuting each variable separately. The grey point cloud shows the null distribution for each variable’s EI, computed after also permuting the validation population group labels 1000 times. Opacity spans from 0% to 100% and is directly inversely proportional to the right-tailed permutation p-value for each variable’s observed EI. If an EI for a variable exceeded 99%, 95%, or 90% of its null values it is demarcated with an asterisk, plus sign, or superscript punctuation, respectively.