Perturbation-Enabled Data Augmentation Improves the Generalizability of Classifiers in Network Neuroscience

Gregory Kiar[1], Alan C. Evans[1], Tristan Glatard[2]

Abstract

Machine learning models are commonly applied to human brain imaging datasets in an effort to associate function or structure with behaviour, health, or other individual attributes. Such models often rely on low-dimensional maps relating brain regions, generated by complex processing pipelines. However, the numerical instabilities inherent to pipelines limits the fidelity of these estimates, and results in bias-rich derivatives serving as inputs to machine learning models. This work seeks to take advantage of numerical instabilities in pipelines by inducing numerical perturbations, ultimately preducing a range of results and reducing the bias in networks used by machine learning models. We found that resampling brain networks across a series of numerically perturbed outcomes led to more consistently generalizable performance in all tested classifiers, preprocessing strategies, and dimensionality reduction techniques when tasked with an age classification task. Importantly, this finding does not hinge on a large number of perturbed networks in order to exhibit improved performance, suggesting that even minimally perturbing a dataset adds meaningful variance which can be captured in the subsequently designed models.

Keywords

Stability — Network Neuroscience — Neuroimaging — Machine Learning — Generalizability

[1] *Montréal Neurological Institute, McGill University, Montréal, QC, Canada*;

[2] *Department of Computer Science and Software Engineering, Concordia University, Montréal, QC, Canada*.

Introduction

The application of machine learning has become commonplace for the identification and characterization of individual biomarkers in neuroimaging[1]. Models may be applied to discriminate between measures of brain structure or function based upon phenotypic variables related to disease[2–4], development[5], or other axes of potential consequence[6,7].

These models often build representations upon processed imaging data, in which 3D or 4D images have been transformed into estimate of structure[8], function[9], or connectivity[10]. However, there is a lack of reliability in these estimates, including variation across analysis team[11], software library[12], operating system[13], and instability in the face of numerical noise[14]. This uncertainty limits the ability of models to learn generalizable relationships among data, and leads to bias-rich predictors. Traditionally, this bias has been reduced through the collection and application of repeated-measurement datasets[15,16], though this requires considerable resources and is not feasible in the context of all clinical populations.

Perturbation methods which inject small amounts of noise throught the execution of a pipeline, such as Monte Carlo Arithmetic (MCA)[17,18], have recently been used to induce instabilities in structural connectome estimation software[19]. Importantly, this technique produces a range of equally plausible results, where no single observation is more or less valid than the others – including those which were left unperturbed. While the impact that sampling from a set of perturbed connectomes may have on learning brain-phenotype relationships[14], there remains potential for leveraging the distribution of results to augment datasets in lieu of increasing sample sizes or performing repeated measurements.

Using an existing collection of MCA-perturbed structural connectomes[20], we trained classifiers on networks sampled from the distribution of results and evaluated their performance relative to using only the unperturbed networks. We evaluate several techniques for resampling the networks, and compare classifiers through their validation performance, the performance on an out-of-sample test set, and the generalizability of their performance across the two. We demonstrate the efficacy of using MCA as a method for dataset augmentation which leads to more robust and generalizable models of brain-phenotype relationships.

Materials & Methods

The objective of this study was to evaluate the impact of aggregating collections of unstable brain networks towards learning robust brain-phenotype relationships. We sampled and aggregated simulated networks within individuals to learn relationships between brain connectivity and individual an trait, in this case age, and compared this to traditional baseline performance on this tasks. We compared aggregation strategies with respect to baseline validation performance, performance out-of-sample, and generalizability.

All developed software and analysis resources for this project have been made available through GitHub at `https://github.com/gkpapers/2020AggregateMCA`.

Dataset

An existing dataset containing Monte Carlo Arithmetic (MCA) perturbed structural human brain networks was used for these experiments[20]. The perturbations introduced for the generation of brain networks in this dataset were at the level of machine-error, simulating expected error over a typical pipeline execution. This dataset contains a single session of data from 100 individuals ($100 \times 1 \times 1$). The derived brain networks were generated with a probabilistic structural connectome estimation pipeline[21] using a fixed random seed, and Monte Carlo Arithmetic (MCA) perturbations were added to all Python-implemented operations throughout the pipeline[17,18]. Each sample was simulated 20 times, resulting in $2,000$ unique graphs. Further information on the processing and curation of this dataset can be found here[14].

This collection enabled the exploration of subsampling and aggregation methods in a typical learning context for neuroimaging[22,23]. Exploring the relationship between the number of simulations and performance further allows for MCA-enabled resampling to be evaluated as a method of dataset augmentation.

As the target for classification, individual-level phenotypic data strongly implicated in brain connectivity was desired. Participant age, which has consistently been shown to have a considerable impact on brain connectivity[24–27], was selected and turned into a binary target by dividing participants into adult (> 18) and non-adult groups (68% adult).

Preprocessing

Prior to being used for this task, the brain networks being were represented as symmetric 83×83 adjacency matrices, sampled upon the Desikan-Killiany-Tourville[28] anatomical parcellation. To reduce redundancy in the data, all edges belonging to the upper-triangle of these matrices were preserved and vectorized, resulting in a feature vector of $3,486$ edges per sample. All samples were preprocessed using one of four standard techniques:

Raw The raw streamline count edge-weight intensities were used as originally calculated.

Log Transform The log10 of edge weights were taken, and edges with 0 weight prior to the transform were reset to 0.

Rank Transform The edges were ranked based on their intensity, with the largest edge having the maximum value. Ties were settled by averaging the rank, and all ranks were finally min-max scaled between 0 and 1.

Z-Score The edge weights were z-scored to have a mean intensity of 0 and unit variance.

Machine Learning Pipelines

The preprocessed connectomes were fed into pipelines consisting of two steps: dimensionality reduction and classification. Dimensionality reduction was applied using one of two methods:

Principal Component Analysis The connectomes were projected into the 20 dimensions of highest variance. The number of components was chosen to capture approximately 90% of the variance present within the dataset.

Feature Agglomeration The number of features in each connectome were reduced by combining edges according to maximum similarity/minimum variance using agglomerative clustering[29]. The number of resulting features was 20, to be consistent with the number of dimensions present after PCA, above.

After dimensionality reduction, samples were fed into one of five distinct classifiers as implemented through scikit learn[30]:

Support Vector Machine The model was fit using a radial basis function (RBF) kernel, L2 penalty, and a balanced regularization parameter to account for uneven class membership.

Logistic Regression A linear solver was used due to the relatively small dataset size. L2 regularization and balanced class weights were used, as above.

K-Nearest Neighbour Class membership was determined using an L2 distance and the nearest 10% of samples, scaling with the number of samples used for training.

Random Forest 100 decision trees were fit using balanced class weights, each splitting the dataset according to a maximum of 4 features per node (corresponding to the rounded square root of 20 total features).

AdaBoost A maximum of 50 decision trees were fit sequentially such that sample weights were iteratively adjusted to prioritize performance on previously incorrectly-classified samples, consistent with[31].

The hyperparameters for all models were refined from their default values to be appropriate for a small and imbalanced dataset. The performance for all pipeline combinations of preprocessing methods, dimensionality reduction techniques, and models using the reference (i.e. unperturbed) executions in the dataset ranged from an F1 score of 0.64–0.875 with a mean of 0.806; this evaluation was performed on a consistent held-out test set which was used for all experiments, as described in a following section. This set of models was chosen as it includes i) well understood standard techniques, ii) both parametric and non-parametric methods, iii) both ensemble and non-ensemble methods, and iv) models which have been commonly deployed for the classification neuroimaging datasets[2–4,6,7,24,32,33].

Dataset Sampling

A chief purpose of this manuscript involves the comparison of various forms of aggregation across equivalently-simulated pipeline outputs. Accordingly, the dataset was resampled prior to dimensionality reduction and classifiers were trained, evaluated, and combined according to the following procedures:

Reference Networks generated without any MCA perturbations were selected for input to the models, serving as a benchmark.

Jackknife The datasets were repeatedly sampled such that a single randomly chosen observation of each unique network was selected (i.e. derived from the same input datum). This resampling was performed 100 times, resulting in the total number of resamplings being $5\times$ larger than the number of unique observations per network, ensuring a broad and overlapping sampling of the datasets.

Median The edgewise median of all observations of the same network were used as the samples for training and evaluation.

Mean Similar to the above, the edgewise mean of all observations for each network were computed and used as input data to the classifiers in both collections.

Consensus A distance-dependent average network[34] was computed across all observations of each network. This data-aware aggregation method, developed for structural brain network analysis, preserves network properties often distorted when computing mean or median networks.

Mega-analysis All observations of each network were used simultaneously for classification, increaseing the effective sample size. Samples were organized such that all observations of the same network only appeared within a single fold for training and evaluation, ensuring double-dipping was avoided.

Meta-analysis Individual classifiers trained across jackknife dataset resamplings, above, were treated as independent models and aggregated into an ensemble classifier. The ensemble was fit using a logistic regression classifier across the outputs of the jackknifed classifiers to learn a relationship between the predicted and true class labels.

The robustness and possible benefit of each subsampling approach was measured by evaluation on a subset of all MCA simulations, including 9 distinct numbers of simulations, ranging from 20 to 2 simulations per sample. Combining the dataset sampling methods, the set of simulations, preprocessing strategies, dimensionality reduction techniques, and classifier models, there were $2,200$ models trained and evaluated.

Training & Evaluation

Prior to training models on the brain networks, 20% of subjects were excluded from each dataset for use as an out-of-sample test dataset for all experiments. With the remaining 80% of subjects, cross validation was performed following a stratified grouped k-fold approach ($k = 5$). In this approach, samples were divided into training and validation sets such that the target variable was proportionally represented on each side of the fold (stratified), conditional upon all observations from the same individual, relevant for the mega-analysis dataset sampling method, falling upon the same side of the fold (grouped). This resulted in 5 fold-trained classifiers per configuration, each trained on 64% of the samples and validated on 16%, prior to each being tested on the remaining 20% of held-out samples. All random processes used in determining the splits used the same seed to remove the effect of random sampling.

Classifiers were primarily evaluated on both the validation and test (out-of-sample) sets using F1 score, a standard measure for evaluating classification performance. The generalizability of predictions was defined as:

$$G = 1 - |F1_{test} - F1_{validation}| \tag{1}$$

where a score of 1 (maximum) indicates the equivalent performance across both the validation and test sets, and a lower score (minimum of 0) indicates inconsistent performance. The absolute change in performance was used in Eq. 1, resulting in a score which penalizes spurious over-performance similarly to under-performance. This is a desired attribute of the measure as all inconsistency, whether due to "luck" or model fit, is undesirable when applying a classifier out-of-sample.

Differences in F1 score and generalizability for perturbed experiments with respect to their reference were used to measure the change in performance between for each dataset sampling technique, and statistical comparisons were made through Wilcoxon Signed-Rank tests.

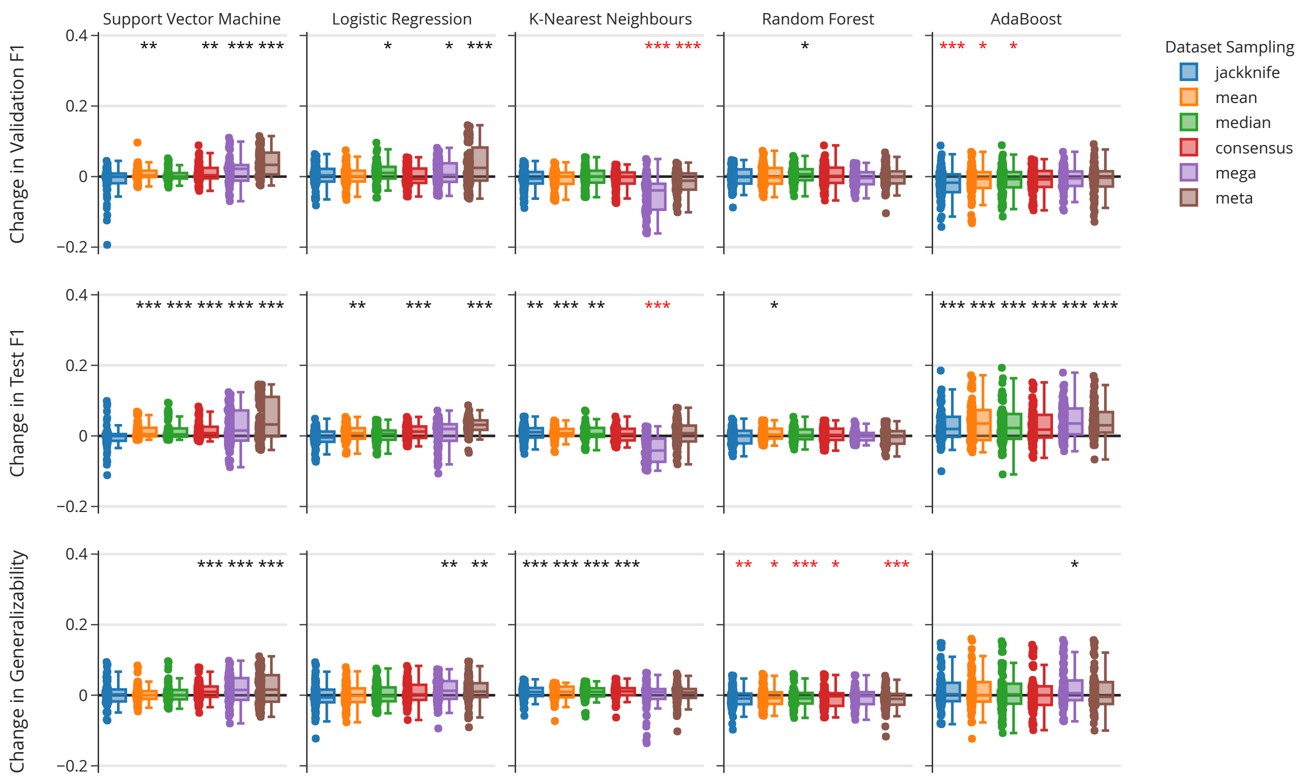

Figure 1. Relative change in classifier performance with respect to classifier type and dataset sampling strategies as measured by change in F1 score on the validation set (top) or test set (middle), as well as the generalizability of performance (bottom). Each star annotation indicates an order of magnitude of statistically significant change, beginning at 0.05 for one star and decreasing from there, with those in black or red indicating an increase or decrease due to resampling, respectively.

Results

The figures and findings presented in this section represent a summary of the complete experiment table which consists of performance measures and metadata for all $2,200$ models tested. The complete performance table alongside the table of significant differences, are made available through the GitHub repository.

Data Resampling Improves Classification

The change in performance for each model and dataset sampling technique is shown in Figure 1. The change in performance was measured as a change in F1 score on the validation set, the change in F1 score on the test set, and the change in overall generalizability, a measure which summarizes the similarity between validation and test performance for a given model.

The support vector machine and logistic regression models improve across each of these three measures for a variety of dataset sampling techniques, meaning that the addition of the MCA-perturbed samples improves the training, testing, and overall generalizability of the classifiers.

Distinctly, k-nearest neighbours (KNN) and AdaBoost classifiers experience minimal change in validation and often see their perfomance decline. However, the improvement of these classifiers on the test set suggests that resampling reduced overfitting in these classifiers. In the case of KNN, this translates to imrpoved generalizability, while in the case of AdaBoost generalizability was largely unchanged, suggesting that the model went from underperforming to overperforming after dataset resampling. The unique decline in performance when using the mega-analytic resampling technique on KNN classifier is suggestive of poor hyperparameterization, as there is a strong relationship between the number of samples in the dataset and the k parameter of the model. At present this parameter was scaled linearly with the number of MCA simulations used, however, it is both possible that an improved scaling function exists or that the model performance degrades with large sample sizes making it a poor model choice given this resampling technique.

The random forest classifiers uniquely did not see a significant change in validation or testing performance across the majority of resampling techniques. However, these classifiers did experience a significant decrease in the generalizability of their performance, meaning that there was a larger discrepancy between training and testing performance in many cases. This distinction from the other models is likely due to the fact that random forest is a simple ensemble technique which takes advantage of training many independent classifiers and samples them to assign final class predictions. It is likely that this approach forms more generalizable predictions generally, and thus the addition of more data does not improve performance further. While AdaBoost is also an ensemble method, the iterative training of models based on sample difficulty allows for the added variance in those samples to play an increasingly central role in the construction of class relationships.

Across all classifier types, it was found that both mega- and meta-analytic approaches outperformed other methods slightly, though this was not statistically significant. Additionally, while certain combinations of preprocessing, dimensionality reduction, and classifiers performed more harmoniously than others, there was no significant relationship between the performance of any single resampling method and preprocessing or dimensionality reduction technique. The above results show that dataset augmentation through MCA-perturbed pipeline outputs may be an effective way to improve the performance and generalizability of non-ensemble classifiers tasked with modeling brain-phenotype relationships, both within and out of sample.

Resampling Leads to Consistent Performance

To beter characterize the the benefit of resampling, the relationship between the magnitude of improvement and the baseline performance of the classifer were further explored (Figure 2). We found that the increase in the generalizability of a classifier was inversely related to the baseline generalizability (Figure 2; top). In other words, the less generalizable a classifier was originally, the more its generalizability improved (significant at $p < 0.05$ for all dataset sampling strategies and classifier other than KNN). There were several situations in which the generalizability of models were noted to decrease, however, though this only occurred for models with high generalizability scores (all > 0.935). Importantly, the relative change in generalizability shifts scores towards a single "mode", suggesting a less biased estimate of the true generalizability of performance on the task, and mitigating both under- and over-performance due to chance.

When exploring the relationship between F1 and generalizability (Figure 2; bottom), it becomes apparent that even for the majority of models which may not have improved performance along both axes, either their generalizability or F1 score is improved. While an ideal classifier would reside in the top-right of the shown plots, the dataset resampling techniques consistently shift the distributions in this direction and improve classifiers along one if not both of these axes. Importantly, the variance in performance across both of these axes is significantly decreased, suggesting that resampling lead to the development of more reliable and reproducible classifiers.

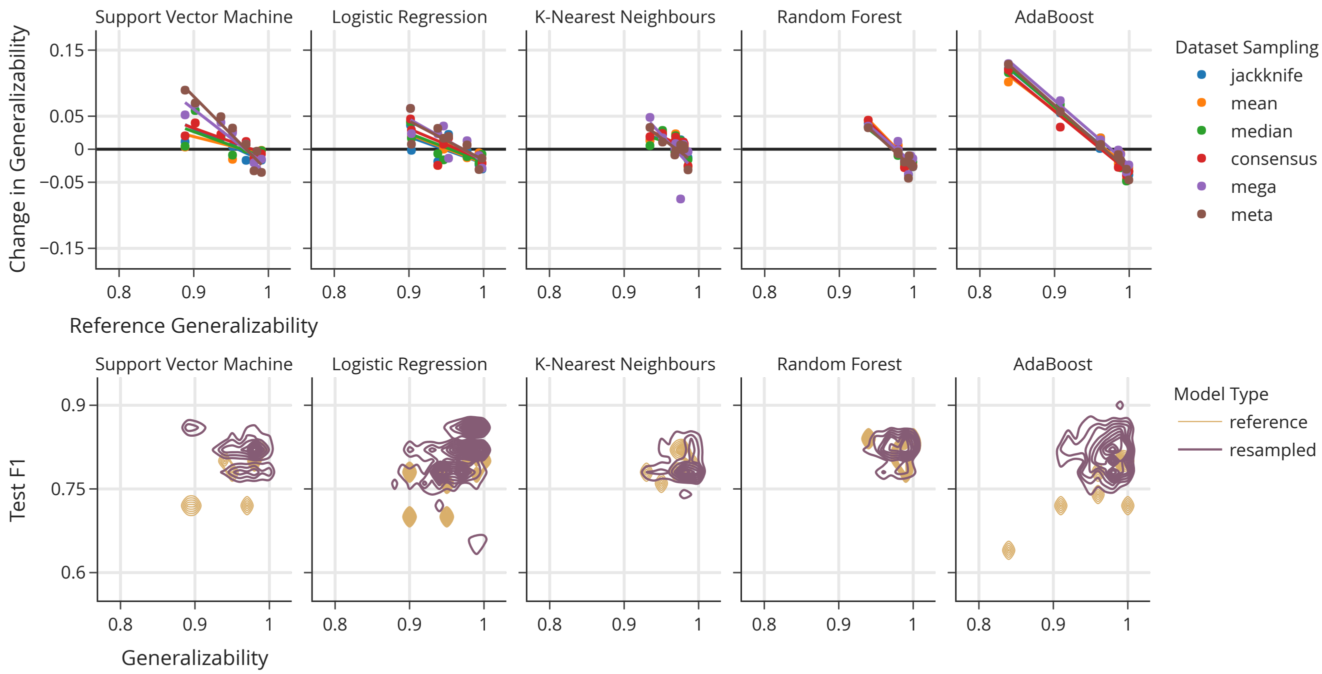

Figure 2. Relationship between generalizability and resampling. Top: change in the generalizability of classifiers with respect to the reference generalizability. Each data point represents the mean change in generalizability for all models using the same preprocessing and dimensionality reduction techniques for a given classifier and dataset sampling strategy. Bottom: contour density distributions of generalizability and F1 scores across all models for both reference and resampled training.

Number of Simulations is Largely Unimpactful

While we previously note an increase in classifier performance due to perturbation-enabled dataset resampling, it was important to explore the relationship between the number of simulated samples and the performance (Figure 3). There was no relationship between the number of independent simulations and performance, as measured by either F1 or generalizability, for all dataset resampling techniques other than mega-analysis. In the case of the mega-analytic approach, however, there was a significant positive relationship between the number of samples used and the generalizability of performance, though there remained no increase in F1 score. The mega-analysis approach is the only approach which changes the number of samples being provided directly to the classifiers, thus mimics an increase in sample size for traditional experiments. While outlying samples may play a small role in many of the proposed forms of resampling, or non-existent in the median case, the mega analytic approach treats all simulations with equal importance as unique samples in the dataset. It is for this reason that the observed relationship here is consistent to what one would expect when increasing the number of samples in their experiment.

Discussion

The numerical perturbation of analytic pipelines provides a unique, data-agnostic, and computationally unintrusive method for dataset augmentation. Using a technique such as MCA, samples can be simulated across an array of controlled executions and used to enrich datasets across a range of plausible results. We demonstrate that this method of dataset augmentation can be used to improve the training, testing, and generalizability of classifiers.

Through the training and evauation of $2,200$ models combining varying strategies for preprocessing, dimensionality

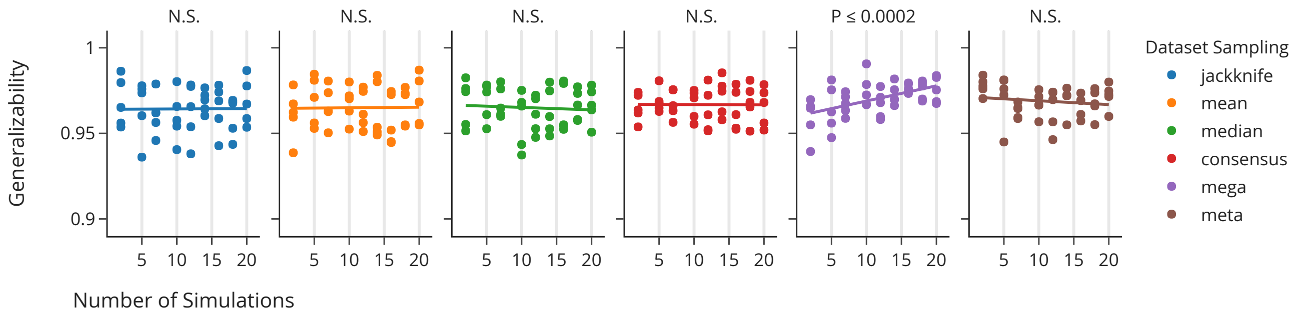

Figure 3. The generalizability of classifiers using each dataset sampling technique with respect to the number of MCA simulations. Each number of simulations was sampled a single time, to avoid artificial skewing of the dataset due to the inclusion of "higher" or "lower" quality samples; a single drawing of each split mimics a true perturbation experiment context.

reduction, classifier, and resampling, we found consistent improvement across all measured axes. Interestingly, while there was a statistically significant improvement in many classifiers, there was no significant difference between the improvement provided by each resampling strategy. This result importantly demonstrates that the added variability in results obtained through MCA is meaningful itself and the most important determination of performance is the inclusion of this variability.

While the non-ensemble methods benefited most obviously from the dataset resampling strategies, where both F1 and generalizability were often improved, the results presented in Figure 2 demonstrate that variability in performance across both of these axes was reduced across all classifier configurations. While a reduction in variability of performance is desirable in itself, this figure also illustrates that the performance of resulting models converges around the more performant models in the case of all classifers.

Although performance was improved by the integration of MCA simulated samples, performance was not significantly related to the number of simulations used in any case other than the mega-analytic resampling strategy. The independence of performance and number of simulations is encouraging, as a key limitation for using Monte Carlo methods is the often extreme computational overhead. The ability to use a small number of simulations and achieve equivalent performance through the majority of resampling techniques allows for significantly better performance without either added data collection and a theoretical doubling the sample processing time. The benefit of increasing the number of simulations in the mega-analytic case could be considered an analog to increasing the sample size of an experiment. While the range of simulations used here demonstrated a consistent improvement in generalizability, there will be a plateau in performance, either at the maximum achievable score or, more likely, before this is reached. Further work is required for characterizing the relationship between the performance of mega-analytic resampling and the number of simulations, and it is likely that this relationship will be domain-specific.

An important limitation of our claim that classifiers with poorer baseline performance benefit more from augmentation is the limit to which that claim remains true. For example, it is unlikely that the trend observed here, with a mean baseline performance of 0.81, would hold across models operating with baseline performance near chance. Characterizing the behaviour of this technique across a range of classification contexts and performances would shed light on whether this technique could be applied globally or if it is limited to making "good" models better.

It is a well understood problem that small sample sizes lead to uncertainty in modeling[35]. This is generally planned for in one of two ways: the collection of vast datasets, as is the case in the UK-BioBank which strives to collect samples from half a million individuals[15]), or the collection of repeated measurements from the selected samples, as is the case in the Consortium of Reliability and Reproducibility which orchestrates multiple centres and scanners, each collecting multiple acquisitions[16]. In either case, the additional data collection by these initiatives is both financially and temporally expensive and leads to unintended confounding effects associated with time of day[36], weather[37], or other off-target variables that are poorly described in the resulting dataset[38]. While the results presented here provide strong evidence in favour of dataset augmentation through numerical perturbations, the improvement from these methods has not be situated relative to additional data acquisition due to the limited sample size of the available perturbed repeated-measures dataset[20]. While it is likely that the methods demonstrated here will work harmoniously with traditional methods of dataset augmentation, it is presently unclear if this technique could serve as a cost-effective replacement for data collection, and exploring that relationship is an exciting avenue for future work.

A common issue in many machine learning contexts is the unbalanced nature of datasets. When using a nearest-neighbour classifier, for instance, a dramatic difference the membership of each group could have significant impact on model hyper-parameters and performance. In contexts where improved sampling is not possible, such as when considering a clinical population, perturbation-augmented datasets could be applied for realistic upsampling of data. In this case, a mega-analytic aggregation strategy could be used in which more simulations would be performed for members of the under-represented class, similar to the balancing of weights applicable to some models. This application is particularly important, as upsampling is often challenging in biological contexts where realistic simulation models are sparse.

Conclusion

This work demonstrates the benefit of augmenting datasets through numerical perturbations. We present an approach which leverages the numerical instability inherent to pipelines for creating more accurate and generalizable classifiers and representations of brain-phenotype relationships. While the approach and results demonstrated here were specifically relevant in the context of brain imaging, the data-agnostic method for inducing perturbations and off-the-shelf machine learning techniques used suggest that this approach may be widely applicable across domains. This work uniquely shows that numerical uncertainty is an asset which can be harnessed to increase the robustness of learned relationships among data.

Data & Code Availability

The perturbed connectomes were publicly available data resource previously produced and made available by the authors[20]. They can be found persistently at `https://doi.org/10.5281/zenodo.4041549`, and are made available through The Canadian Open Neuroscience Platform (`https://portal.conp.ca/search`, search term "Kiar"). All software developed for processing or evaluation is publicly available on GitHub at `https://github.com/gkpapers/2020AggregateMCA`. Experiments were launched on Compute Canada's HPC cluster environment.

Author Contributions

GK was responsible for the experimental design, data processing, analysis, interpretation, and the majority of writing. All authors contributed to the revision of the manuscript. TG and ACE contributed to experimental design, analysis, interpretation. The authors declare no competing interests for this work. Correspondence and requests for materials should be addressed to Gregory Kiar at `gregory.kiar@mail.mcgill.ca`.

Acknowledgments

This research was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (award no. CGSD3-519497-2018). This work was also supported in part by funding provided by Brain Canada, in partnership with Health Canada, for the Canadian Open Neuroscience Platform initiative.

References

[1] C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager, "Building better biomarkers: brain models in translational neuroimaging," *Nature neuroscience*, vol. 20, no. 3, p. 365, 2017.

[2] N. A. Crossley, A. Mechelli, J. Scott, F. Carletti, P. T. Fox, P. McGuire, and E. T. Bullmore, "The hubs of the human connectome are generally implicated in the anatomy of brain disorders," *Brain*, vol. 137, no. Pt 8, pp. 2382–2395, Aug. 2014.

[3] S. Payabvash, E. M. Palacios, J. P. Owen, M. B. Wang, T. Tavassoli, M. Gerdes, A. Brandes-Aitken, D. Cuneo, E. J. Marco, and P. Mukherjee, "White matter connectome edge density in children with autism spectrum disorders: Potential imaging biomarkers using Machine-Learning models," *Brain Connect.*, vol. 9, no. 2, pp. 209–220, Mar. 2019.

[4] E. Tolan and Z. Isik, "Graph theory based classification of brain connectivity network for autism spectrum disorder," in *Bioinformatics and Biomedical Engineering*. Springer International Publishing, 2018, pp. 520–530.

[5] M. Zhang, C. Desrosiers, Y. Guo, B. Khundrakpam, N. Al-Sharif, G. Kiar, P. Valdes-Sosa, J.-B. Poline, and A. Evans, "Brain status modeling with non-negative projective dictionary learning," *Neuroimage*, p. 116226, Oct. 2019.

[6] X. Zhu, X. Du, M. Kerich, F. W. Lohoff, and R. Momenan, "Random forest based classification of alcohol dependence patients and healthy controls using resting state MRI," *Neurosci. Lett.*, vol. 676, pp. 27–33, May 2018.

[7] B.-Y. Park, J. Seo, J. Yi, and H. Park, "Structural and functional brain connectivity of people with obesity and prediction of body mass index using connectivity," *PLoS One*, vol. 10, no. 11, p. e0141376, Nov. 2015.

[8] B. S. Wade, S. H. Joshi, B. A. Gutman, and P. M. Thompson, "Machine learning on high dimensional shape data from subcortical brain surfaces: A comparison of feature selection and classification methods," *Pattern Recognition*, vol. 63, pp. 731–739, 2017.

[9] S. Weis, K. R. Patil, F. Hoffstaedter, A. Nostro, B. T. Yeo, and S. B. Eickhoff, "Sex classification by resting state brain connectivity," *Cerebral cortex*, vol. 30, no. 2, pp. 824–835, 2020.

[10] B. C. Munsell, C.-Y. Wee, S. S. Keller, B. Weber, C. Elger, L. A. T. da Silva, T. Nesland, M. Styner, D. Shen, and L. Bonilha, "Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data," *Neuroimage*, vol. 118, pp. 219–230, 2015.

[11] R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock *et al.*, "Variability in the analysis of a single neuroimaging dataset by many teams," *Nature*, pp. 1–7, 2020.

[12] A. Bowring, C. Maumet, and T. E. Nichols, "Exploring the impact of analysis software on task fmri results," *Human brain mapping*, vol. 40, no. 11, pp. 3362–3384, 2019.

[13] T. Glatard, L. B. Lewis, R. Ferreira da Silva, R. Adalat, N. Beck, C. Lepage, P. Rioux, M.-E. Rousseau, T. Sherif, E. Deelman, N. Khalili-Mahani, and A. C. Evans, "Reproducibility of neuroimaging analyses across operating systems," *Front. Neuroinform.*, vol. 9, p. 12, Apr. 2015.

[14] G. Kiar, Y. Chatelain, P. de Oliveira Castro, E. Petit, and others, "Numerical instabilities in analytical pipelines lead to large and meaningful variability in brain networks," *bioRxiv*, 2020.

[15] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *Plos med*, vol. 12, no. 3, p. e1001779, 2015.

[16] X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. C. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos *et al.*, "An open science resource for establishing reliability and reproducibility in functional connectomics," *Scientific data*, vol. 1, no. 1, pp. 1–13, 2014.

[17] D. S. Parker, *Monte Carlo Arithmetic: exploiting randomness in floating-point arithmetic*. University of California (Los Angeles). Computer Science Department, 1997.

[18] C. Denis, P. de Oliveira Castro, and E. Petit, "Verificarlo: Checking floating point accuracy through monte carlo arithmetic," *2016 IEEE 23nd Symposium on Computer Arithmetic (ARITH)*, 2016.

[19] G. Kiar, P. de Oliveira Castro, P. Rioux, E. Petit, S. T. Brown, A. C. Evans, and T. Glatard, "Comparing perturbation models for evaluating stability of neuroimaging pipelines," *The International Journal of High Performance Computing Applications*, 2020.

[20] G. Kiar, "Numerically perturbed structural connectomes from 100 individuals in the NKI rockland dataset," Apr. 2020.

[21] E. Garyfallidis, M. Brett, B. Amirbekian, A. Rokem, S. van der Walt, M. Descoteaux, I. Nimmo-Smith, and Dipy Contributors, "Dipy, a library for the analysis of diffusion MRI data," *Front. Neuroinform.*, vol. 8, p. 8, Feb. 2014.

[22] S. I. Dimitriadis, M. Drakesmith, S. Bells, G. D. Parker, D. E. Linden, and D. K. Jones, "Improving the reliability of network metrics in structural brain networks by integrating different network weighting strategies into a single graph," 2017.

[23] C. R. Buchanan, C. R. Pernet, K. J. Gorgolewski, A. J. Storkey, and M. E. Bastin, "Test–retest reliability of structural brain networks from diffusion MRI," *Neuroimage*, vol. 86, pp. 231–243, Feb. 2014.

[24] T. B. Meier, A. S. Desphande, S. Vergun, V. A. Nair, J. Song, B. B. Biswal, M. E. Meyerand, R. M. Birn, and V. Prabhakaran, "Support vector machine classification and characterization of age-related reorganization of functional brain networks," *Neuroimage*, vol. 60, no. 1, pp. 601–613, Mar. 2012.

[25] K. Wu, Y. Taki, K. Sato, S. Kinomura, R. Goto, K. Okada, R. Kawashima, Y. He, A. C. Evans, and H. Fukuda, "Age-related changes in topological organization of structural brain networks in healthy individuals," *Hum. Brain Mapp.*, vol. 33, no. 3, pp. 552–568, Mar. 2012.

[26] S. Y. Bookheimer, D. H. Salat, M. Terpstra, B. M. Ances, D. M. Barch, R. L. Buckner, G. C. Burgess, S. W. Curtiss, M. Diaz-Santos, J. S. Elam, B. Fischl, D. N. Greve, H. A. Hagy, M. P. Harms, O. M. Hatch, T. Hedden, C. Hodge, K. C. Japardi, T. P. Kuhn, T. K. Ly, S. M. Smith, L. H. Somerville, K. Uğurbil, A. van der Kouwe, D. Van Essen, R. P. Woods, and E. Yacoub, "The lifespan human connectome project in aging: An overview," *Neuroimage*, vol. 185, pp. 335–348, Jan. 2019.

[27] T. Zhao, M. Cao, H. Niu, X.-N. Zuo, A. Evans, Y. He, Q. Dong, and N. Shu, "Age-related changes in the topological organization of the white matter structural connectome across the human lifespan," *Hum. Brain Mapp.*, vol. 36, no. 10, pp. 3777–3792, 2015.

[28] A. Klein and J. Tourville, "101 labeled brain images and a consistent human cortical labeling protocol," *Front. Neurosci.*, vol. 6, p. 171, Dec. 2012.

[29] J. H. Ward, "Hierarchical grouping to optimize an objective function," pp. 236–244, 1963.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and Others, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[31] Y. Freund and R. E. Schapire, "A Decision-Theoretic generalization of On-Line learning and an application to boosting," *J. Comput. System Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[32] B. Tunç, B. Solmaz, D. Parker, T. D. Satterthwaite, M. A. Elliott, M. E. Calkins, K. Ruparel, R. E. Gur, R. C. Gur, and R. Verma, "Establishing a link between sex-related differences in the structural connectome and behaviour," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 371, no. 1688, p. 20150111, Feb. 2016.

[33] D. R. Nayak, R. Dash, and B. Majhi, "Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests," *Neurocomputing*, vol. 177, pp. 188–197, Feb. 2016.

[34] R. F. Betzel, A. Griffa, P. Hagmann, and B. Misic, "Distance-dependent consistency thresholds for generating group-representative structural brain networks," *bioRxiv*, 2018.

[35] G. Varoquaux, "Cross-validation failure: small sample sizes lead to large error bars," *Neuroimage*, vol. 180, pp. 68–77, 2018.

[36] G. Vandewalle, S. N. Archer, C. Wuillaume, E. Balteau, C. Degueldre, A. Luxen, P. Maquet, and D.-J. Dijk, "Functional magnetic resonance imaging-assessed brain responses during an executive task depend on interaction of sleep homeostasis, circadian phase, and per3 genotype," *Journal of Neuroscience*, vol. 29, no. 25, pp. 7948–7956, 2009.

[37] X. Di, M. Wolfer, S. Kühn, Z. Zhang, and B. B. Biswal, "Estimations of the weather effects on brain functions using functional mri: A cautionary tale," 2019.

[38] L. Chaddock, K. I. Erickson, R. S. Prakash, J. S. Kim, M. W. Voss, M. VanPatter, M. B. Pontifex, L. B. Raine, A. Konkel, C. H. Hillman *et al.*, "A neuroimaging investigation of the association between aerobic fitness, hippocampal volume, and memory performance in preadolescent children," *Brain research*, vol. 1358, pp. 172–183, 2010.