Predicting post-Stroke Cognitive Impairments from Lesion Topography using Machine Learning

Muhammad Hasnain Mamdani

School of Computer Science
McGill University, Montreal

February, 2021

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Science.

Abstract

Background: Stroke is the fourth and fifth leading cause of death in Canada and the United States. Survivors of stroke live with mild to severe life-long impairments. Early rehabilitation can improve long-term outcomes of stroke patients and improve their quality of life. Accurate prediction of post-stroke cognitive impairments at an individual patient level may aid the development of personalized treatments and intervention strategies.

Methods: We applied and benchmarked machine learning methods on a relatively large stroke dataset (n=1401) to predict cognitive outcomes from lesion topography. The dataset included MRIs (Structural axial T1, T2-weighted spin echo, DWI and FLAIR sequence) of ischemic stroke patients carried out within 7 days from the onset of stroke and their neuropsychological assessments including measures for global cognition, language, memory, visuospatial functioning, information processing speed and executive functioning at 3 months. Three approaches to analyzing brain-behavior relationships from a predictive analytics standpoint were explored and compared in terms of out-of-sample prediction performance of post-stroke cognitive functions based on 5-fold nested cross-validation: 1) multi-outcome models vs single-outcome models; 2) non-linear models vs linear models; and 3) data augmentation (Mixup).

Results: The out-of-sample coefficient of determination (r-square) values in all approaches are generally low and inconsistent across cross-validation folds indicating poor predictive performance. However, we see that: 1) joint modeling of interrelated cognitive functions exhibits potential to perform more accurate predictions in the domains of global cognition and language; 2) non-linear models could potentially be exploited to improve individualized predictions in the domains of language and memory; and 3) it is not easy to exploit artificial samples generated by Mixup to improve the predictive performance of cognitive functions post-stroke.

Conclusion: Prediction of single patient outcomes from lesion topography is a difficult task with the quality and quantity of neuroimaging data currently available for stroke. This work highlights the challenges and provides useful directions to future research in lesion-behavior mapping.

Résumé

Contexte: L'AVC est la quatrième et la cinquième cause de décès en importance au Canada et aux États-Unis. Les survivants d'un AVC vivent avec des déficiences légères à sévères à vie. Une rééducation précoce peut améliorer les résultats à long terme des patients victimes d'un AVC et améliorer leur qualité de vie. Une prédiction précise des déficiences cognitives post-AVC au niveau d'un patient individuel peut aider au développement de traitements et de stratégies d'intervention personnalisés.

Méthodes: Nous avons appliqué et comparé des méthodes d'apprentissage automatique sur un échantillon plutôt large de données d'AVC (n = 1401) pour prédire les résultats cognitifs à partir de la topographie des lésions. La banque de données comprenait des images IRM (structurelles axiales T1, spin écho pondérées en T2, DWI et séquence FLAIR) de patients victimes d'un AVC ischémique réalisées dans les 7 jours suivant le début de l'AVC et leurs évaluations neuropsychologiques, y compris des mesures de la cognition globale, du langage, de la mémoire, du fonctionnement visuospatial , dela vitesse de traitement de l'information et de la fonction exécutive à 3 mois. Trois approches pour analyser les relations cerveau-comportement du point de vue de l'analyse prédictive ont été explorées et comparées en termes de performances de prédiction hors échantillon des fonctions cognitives post-AVC, basées sur une validation croisée imbriquée 5 fois: 1) modèles multi-variés vs modèles univariés; 2) modèles non linéaires vs modèles linéaires 3) augmentation des données (Mixup).

Résultats: Les valeurs du coefficient de détermination hors échantillon (r-carré) dans toutes les approches sont généralement faibles et incohérentes entre les plis de validation croisée, ce qui indique une performance prédictive médiocre. Cependant, nous voyons que: 1) la modélisation conjointe de fonctions cognitives interdépendantes présente le potentiel d'effectuer des prédictions plus précises dans les domaines de la cognition globale et du langage; 2) les modèles non

linéaires pourraient potentiellement être exploités pour améliorer les prédictions individualisées dans les domaines du langage et de la mémoire; et 3) il n'est pas facile d'exploiter des échantillons artificiels générés par Mixup pour améliorer les performances prédictives des fonctions cognitives après un AVC.

Conclusion: La prédiction des résultats d'un seul patient à partir de la topographie des lésions est une tâche difficile compte tenu de la qualité et de la quantité de données de neuroimagerie actuellement disponibles pour les AVC. Ce travail met en évidence les défis et fournit des orientations utiles pour les recherches futures sur la cartographie lésion-comportement.

Acknowledgements

I would like to thank my supervisor, Professor Danilo Bzdok, for taking me as one of his first students in his new interdisciplinary lab in Montreal on joining McGill University. He patiently transferred his knowledge and expertise to me on many topics related to machine learning, statistics, and neuroscience and connected me with notable experts in the areas of machine learning and imaging neuroscience. I would like to express gratitude for his financial contributions to support this research. I also wish to express thanks to my second supervisor Professor Blake Richards for his intellectual contribution towards understanding the limitations of the link between brain imaging measurements and cognitive processes. Furthermore, I am thankful to Professor Anthony Randal McIntosh for his valuable feedback on the initial version of this thesis.

I would like to thank our collaborators Nick and Hugo from Utrecht University in the Netherlands who kindly made accessible to us one of the largest databases of stroke patients to date. This includes lesion segmented structural MRIs and neuropsychological assessment scores of 1401 subjects. They had been very kind in providing interpretations and useful insights on the dataset and clinical aspects of the stroke.

2020 has been an unusual year with the COVID-19 pandemic induced social restrictions. I consider myself privileged to be connected with my parents, sisters, family, and close friends, who boosted my morale during times of global crisis. I am grateful for their social support, in-person and virtual, which helped me advance my research goals.

Many thanks to my colleagues in the lab Hannah, Emile and Nahiyan. Most of us were new to the lab and spent little time together in person in the lab on a daily basis before the pandemic started. But in those few months, we exchanged interesting ideas, engaged in intellectual discussions and developed

lasting friendships. They were always there to support me throughout the journey of this project, in-person before the pandemic and virtual after.

Contribution of Authors

The research conducted in this thesis including the study design, implementation of machine learning methods, results and discussion as well as the writing of the thesis are the original work of the student with guidance and continuous feedback provided by the supervisors.

Contents

List of Figures

1

Introduction

Stroke is the fourth leading cause of death in Canada in 2019. In Canada, more than 50,000 people die of stroke every year [14]. Stroke is typically a result of chronic disease of the arteries and their impairment leads to the destruction of brain tissue in an acute disease episode. Blood vessels block causing ischemic injury to brain tissue, or rupture (hemorrhagic stroke) resulting in bleeding in the brain. The effect of stroke can be seen in motor function (e.g. paralysis) or in cognitive impairments (e.g. memory loss). The magnitude of these impairments depends on several factors, including the topography of ischemia and the amount of brain tissue affected. Stroke lesions do not only result in motor impairments but also a range of cognitive impairments from mild to severely disabling symptoms [34].

Stroke patients undergo diagnostic imaging such as Computed Tomography (CT) scan or Magnetic Resonance Imaging (MRI) shortly after the episode of stroke allowing the clinicians to examine the structural damage in the brain. The measurement of clinical impairment is done through a physical examination for motor impairment and neurophysiological assessment for cognitive functioning. Clinicians employ a suite of standardized tests to measure patients' global cognition, language performance, memory function, visuospatial function, motor function, attention, information processing speed and executive function. This battery of behavioral tests is conducted at various time

points after stroke.

The impairments as a result of a stroke can persist and affect the quality of life [23]. Early rehabilitation can improve long term outcomes of physical and cognitive domains [19], which underlines the potential of early prediction of individual-patient level cognitive outcomes. Accurate prediction of cognitive impairments would enable the development of personalized treatments and rehabilitation plans and provide valuable insights to tailor intervention strategies. Identifying patients that may develop serious cognitive impairments post-stroke and offer them appropriate and timely rehabilitation is the goal of prediction based studies [33].

In this study, we aim to predict the cognitive outcomes of patients based on the spatial topography of lesions. The cognitive outcomes include the patient's global cognition, language performance, memory function, visuospatial function, information processing speed and executive function as measured by a battery of standard neuropsychological tests (cf. Methods). We investigate various approaches to analyzing brain-behavior relationships from a predictive analytics standpoint: multi-outcome modeling, non-linear modeling and data augmentation via Mixup [65]. We introduce these concepts below and shed light on the motivation behind these approaches.

1.1 Multi-Outcome Modeling

Stroke patients typically do not experience impairment in only one cognitive domain (e.g. memory, language, information processing speed, etc.), but many at the same time. Figure 1.1 shows that impairments in various cognitive functions co-occur in our database of stroke patients.

We want to explicitly acknowledge the mutual relatedness of various cognitive functions and model this general property of stroke in our modeling. To do so, we

use multi-output (a.k.a. multi-task) learning models [15] [5] [13] to jointly predict cognitive function scores from the lesion topography based on structural brain MRIs and known brain atlases.

Multi-output learning models are those that simultaneously model (and predict) multiple target variables. Unlike single-outcome models that associate a set of input variables with one output variable, multi-outcome models associate a set of input variables to a set of output variables.

With multi-output modeling, we aimed to analyze whether various cognitive deficits share similar underlying disease processes and therefore exploit the existence of any shared patterns across various clinical dimensions. In essence, we would like to examine whether joint modeling of an array of cognitive functions improves predictive performance [45].

1.2 Non-Linear Modeling

It is known that the human brain neural network is a highly non-linear and complex system. Our aim is to try to capture (some part of) the non-linearity captured in brain imaging measurements, if so, in our analysis to better explain cognitive functions.

We define linear models as those where observational data are modeled by a function that is a linear combination of the input variables, for e.g. linear regression. In contrast, models in which the observational data are not modeled by a linear combination of the input variables are considered non-linear, for e.g. logistic regression.

With non-linear modeling, we want to explore the hypothesis that lesion volumes in known brain regions interact non-linearly to explain cognitive scores. The motive behind non-linear modeling is to seek quantitative evidence of how a more elaborate interplay of spatially distributed parts of the brain impacts cognitive functioning. For

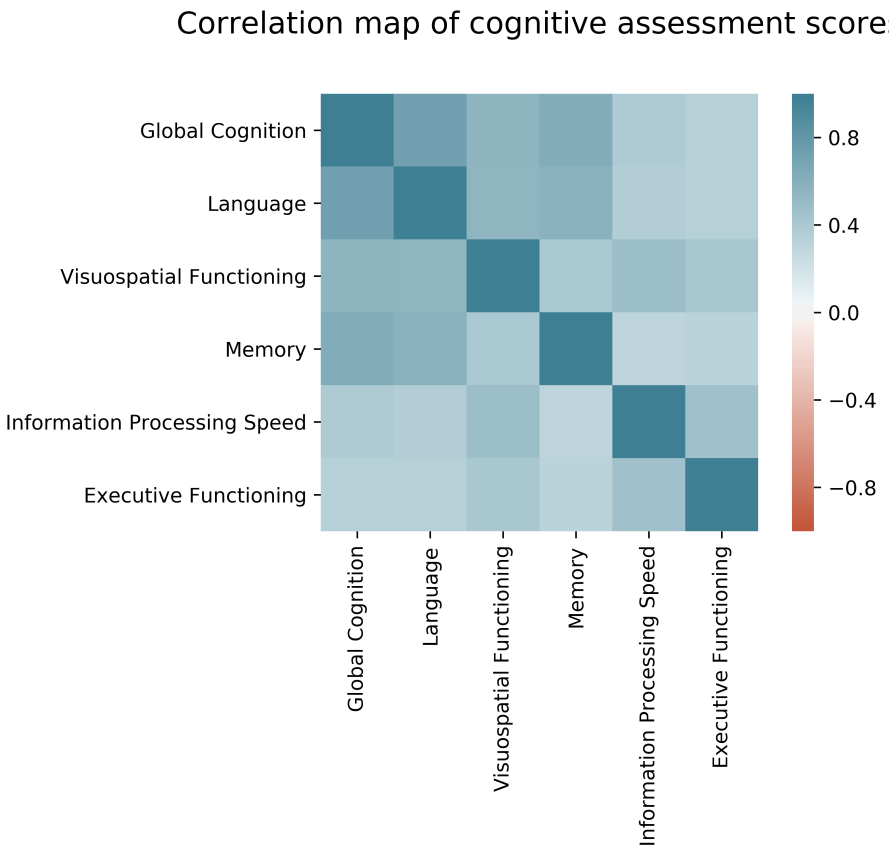

Figure 1.1: Pearson's correlation coefficient between the six cognitive assessment scores of 1401 stroke patients. The values ranging from 0.30-0.72 indicate that the cognitive deficits co-occur in multiple domains in patients suffering from a stroke. The highly correlated scores hint at the presence of shared latent factor(s).

e.g., is there a non-linear interaction between the Broca's region on the left and the angular gyrus on the right that could be exploited by non-linear modeling to predict language abilities?

Eventually, with more ambitious non-linear models, we seek more accurate predictions on a single subject level. We would like to examine whether non-linear models could predict behavior better on average compared to simple linear models, thereby assisting with better diagnosis and treatments of stroke patients.

1.3 Mixup (Data Augmentation)

Modern machine learning models show great promise in predictive capabilities but often require a huge sample size for training. The neuroimaging and behavioral dataset of stroke patients that one researcher has access to is often limited in sample size, rarely reaching four digits. In theory, stroke data is abundant. In the US alone, about 795,000 people experience a stroke every year [56]. However, challenges exist to coordinate and share medical data between hospitals and between regions [3].

The small size of the dataset limit the ability of otherwise powerful machine learning methods to discover quality brain-behavior patterns. In this study, we use Mixup, a modern data augmentation approach to address the problem of scarcity of data. We examine whether increasing the sample size by *creating* new observations that are plausible variants of the original data could better explain and predict cognitive functions from lesion topography. Artificially constructing useful samples that are otherwise hard to obtain would be yet another utility to the stroke community.

Mixup creates additional samples of data by interpolating two true samples from the original dataset taken at random. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be two input vectors and $\mathbf{y}_i$ and $\mathbf{y}_j$ their corresponding output vectors in the dataset. Mixup generates virtual samples $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ as follows:

$$\begin{aligned} \hat{\mathbf{x}} &= \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \\ \hat{\mathbf{y}} &= \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j, \end{aligned} \tag{1.1}$$

where $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}_j, \mathbf{y}_j)$ are two samples drawn at random from the training data and $\lambda \in [0, 1]$. To illustrate this, let's say two samples from the database are used by Mixup to create a virtual sample: one where the patient has a stroke lesion in region A and impaired language ability, and second where the patient has a stroke

lesion in region B and impaired memory performance. The artificial sample created based on these given samples will be where a patient has lesions in both region A and region B and impairments in both language and memory function, the extent of which is determined by λ.

An important characteristic of Mixup is that, unlike many other data augmentation approaches, Mixup generates new samples of not only the input variables but also the output variables. Mixup extends the data by incorporating an inductive bias that linear interpolations of input vectors lead to linear interpolations of associated output variables. This could be useful in exploring the hypothesis that lesion volumes in known brain regions are linearly associated with cognitive deficits.

In this study, we have stroke data from a broad population and multiple cohorts. Therefore, with Mixup, we hope to generate additional samples that are close to the true distribution of stroke patients. Eventually, with a larger dataset, we aim to potentially exploit more ambitious and data-hungry models to better predict post-stroke cognitive outcomes.

2

Literature Review

Cognitive neuroscience has progressed in the last century through the study of patients with brain lesions, particularly stroke patients. What lesion studies of stroke patients have brought to knowledge in neuroscience is debated among experts, with various views being brought upon their value and their inherent challenges [3].

Lesion analysis is a classic approach to study brain structures and their functioning. The type and extent of a stroke lesion and its evolution can give clinicians insights about the impairments and their potential of resolution over time. Lesion-symptom mapping (LSM) studies have been applied to various clinical conditions such as motor function impairment [49], aphasia [66] [63] [29], spatial neglect [60] and other cognitive impairments (memory, executive functioning, etc.) [21] [67]. Studies have made various assumptions in lesion-symptom mapping (LSM): 1) a region of the brain and a specific behavior are linked; 2) the link between a brain region and cognitive process exists for all functional domains; 3) a specific voxel can be linked to a behavioral deficit through statistical significant association. These assumptions are now being revisited [20]

Structural imaging-based lesion-symptom mapping (LSM) with MRI is a popular approach used in studies looking at brain-behavior relationships. Other imaging modalities like CT-Scan, Positron-Emission Tomography (PET) and now functional MRI [45] can be complementary to structural MRIs [57] in LSM studies.

Lesion symptom mapping is generally performed by providing a statistical model brain imaging data as an input and a certain behavior score as an output. Studies have evolved from using only lesion volume as an input feature, to integrating the location of the lesion for modeling cognitive outcomes. The input variables are either individual voxels of the brain image or some pre-processed form of them. Voxel-based lesion-symptom mapping (VLSM) was the first evolution in traditional LSM, which explored impact of one voxel at a time [7]. Assessing the brain functions on a voxel-by-voxel basis allows identifying brain locations that are related to cognitive deficits. The significance of the behavioral difference (continuous variable) is analyzed using parametric and non-parametric methods.

This mass univariate approach, however, is less meaningful biologically as it does not consider the interaction of neighboring or spatially distant regions of the brain. Hence, studies started adapting 'multivariate' approaches where all brain regions are provided concurrently as input to a single model. A structural MRI of 1 mm^3 resolution contains approximately 2 million voxels making it impractical for most multivariate models to directly use all brain regions in the raw voxel form at once with a sample size of a mere few hundred typical in LSM studies. To overcome this curse of dimensionality, two main methods (and variations of them) have generally been used to reduce the number of input variables: aggregating voxels in regions of interest (ROI) based on standard brain atlases [33] [25] [67] and Principal Component Analysis (PCA) [29] [38] [54]. Multivariate models enable the detection of statistically significant anatomical networks (topography) that contribute to a cognitive deficit. These methods enable the search for shared patterns that are linked to a cognitive domain [3]. There is a growing body of literature that uses such 'inference-based' methods [25] [46] [21] [54] [29].

Studies have highlighted the association of white matter regions with cognitive outcomes. Corbetta et al. [21] used multivariate modeling to measure the proportion

of behavioral variance that could be imputed to structural lesions. They observed that impairment can be explained by some clusters of cognitive deficits underlying multiple functions. The observed cluster was located in the subcortical as well as white matter regions. They underscored the need for better models to associate cognitive impairments with white matter damage. Ramsey et al., in a study [46], examined patterns and variability of post-stroke recovery in multiple behavioral domains (44 neuropsychological tests) and showed that white matter damage impacts various cognitive functions. They reinstated the importance of white matter lesions in brain-behavior relationships, just like cortical lesions that have classically been studied. Yourganov et al. in a predictive study [63] of types of aphasia, explored five atlases and shows that the Support Vector Machine (SVM) performs the most accurate classification when provided with the combination of gray and white matter atlases.

Multivariate LSM studies have explored the importance of the location of the infarct along with its volume in association with behavioral deficits [18] [62] [39] [67]. Wu et al. [62] used stroke lesion topography to determine the significance of infarct location on acute (early) ischemic stroke severity and long-term modified Rankin scale score. Their work gave inferential insights on the relationship between brain regions and functional outcomes. They concluded on the importance of including the location of the stroke, as well as volume and some socio-demographic characteristics in future predictive modeling studies. They also discussed that the Rankin score is too global and specific cognitive outcomes should be used.

Zhao et al. in a study [67] showed that global cognition (including memory, language, visuospatial and executive functions) and infarct location are linked. They used multivariate analyses with regions of interest (ROI) using support vector regression (SVR) to explore the brain-behavior relationship. They confirmed the importance of various cortical and sub-cortical regions in relation to specific cognitive domains.

Both studies [62] [67] suggested that future studies should test these models on independent cohorts that are not used in the training of the model, stressing the importance of developing the predictive studies regime. Predictive studies are centered on clinical endpoint prediction from a precision medicine perspective, whereas inference-based studies focus primarily on the mechanistic understanding of brain regions that contribute to a disease [9]. The LSM methods are evolving to trying to predict outcomes from early imaging of stroke patients to be able to potentially provide clinicians with tools to adjust therapy for individual patients.

In the early studies of predictive modeling, Hope et al. proposed Gaussian Process Regression (GPR), an algorithm integrating patient's lesion and demographic information, to predict the speech recovery at different time points post-stroke in a new set of patients (out-of-sample testing) [33]. This early work based on 270 patients demonstrated that lesion volume and location could help predict behavioral outcomes over time, thereby opening pathways for prediction of prognoses of individual patients.

In another early prediction based study, Zhang et al. worked with aphasia patients exploring a multivariate non-linear model (SVR-LSM) for lesion mapping capable of integrating voxel connections. Their work highlighted important brain regions linked to aphasia, but their predictive accuracy was low, showing the difficulty in language function prediction. Their mean prediction performance (R^2) was 0.10 for semantic error (SE) and 0.11 for phonological error (PE) [66].

Rondina et al. did a predictive study [49] to identify patients that will have a better recovery of motor impairment of the upper limb using as input T1-weighted structural brain scans. They used support vector machine approach with voxel-wise lesion likelihood values to show that they can classify patients with better recovery of prognosis. This work showed the importance of various structures of the brain associated with motor function of the upper limb and indicated that the prediction

methodology could in the future be refined so to stratify patients for rehabilitation trials.

Munsch et al. explored the importance of stroke location in a predictive objective [39]. Authors used multivariate models to predict general functional outcome (Rankin scale) as well as cognitive outcomes (MoCA). They first used VLSM to find eloquent regions; then they developed two models: the first one using the classic inputs (National Institutes of Health Stroke Scale (NIHSS) score, age and infarct volume, and the second model added stroke location. They showed that including stroke location significantly improves predictive results (area under the curve increased from 0.697–0.771; difference=0.073; 95% confidence interval, 0.008–0.155). Results were replicated in out-of-sample data. Authors concluded that that stroke location is of importance for the prediction of MoCA cognitive outcomes at three months.

Ramsey et al. also looked at the impact of lesion location on chronic impairments using eleven principal components that explain 60% of the variance [46]. They also examined the prediction of various chronic scores and found that some models are able to significantly better explain language recovery (13% variance explained, $P < 0.001$), motor (4% variance explained, $P < 0.05$) and attention scores (14% variance explained, $P < 0.05$, but not memory function. They observed that the percentage of variance explained across domains is small compared to the impact of acute impairment score in predicting chronic outcomes.

Aben et al. [1] used diffusion tensor imaging-based measures of brain connectivity to predict one-year cognitive recovery (also integrating other variables such as patient characteristics and stroke severity). The PROCRAS (Prediction of Cognitive Recovery After Stroke) study showed that strategic areas of the brain network in the white matter, which they call “hubs”, can be identified: these hubs, when added to lesion topography and size, can predict recovery in cognitive domains. They proposed a

lesion impact score that would reflect damage to these network hubs. A lower lesion impact score was an independent predictor of cognitive recovery 1 year after stroke (odds ratio=0.434 [0.193–0.978]; P=0.044).

Moulton et al. [38] suggested that lesioned voxel could be characterized by continuous variables that capture the severity of infarct (instead of usual binary segmentation). For the task of prediction of long term cognitive outcomes (Rankin Score: good (mRS ≤ 2) and poor (mRS > 2)), SVM classifiers showed a median [IQR] accuracy of 82.8 [79.3–86.2]% with axial diffusivity maps compared to an accuracy of 76.7 [73.3–82.8]% with lesion segmentations. Their work illustrate that raw continuous information provided in MRIs could be beneficially exploited to predict post-stroke cognitive outcomes.

Chauhan et al. [17] applied newer methods using convolutional neural networks (CNNs) and compared results to PCA, Ridge and SVR to predict the severity of language disorder in stroke patients from structural brain MRIs. A novel combination of CNN and classical Ridge regression (hybrid method) showed the best predictive performance in most cases. CNNs have the advantage of operating directly on raw images. However, their utility is limited in the neuroimaging domain due to CNN's inherent translational invariance property (that does not take into account the specific location in space of image features) as this property does not necessarily hold true in the context of brain functioning.

In most predictive studies [54] [67] [29] [17] [60], out-of-sample validation has been done, but the methods to do so are heterogeneous. Only two studies explicitly did a nested cross-validation to calculate the out-of-sample prediction performance [63] [38]. Out-of-sample performance is crucial as it represents the applicability of a model to new subjects that are not part of the data used in the fitting/training of the model.

In all these studies, the major bottleneck is limited patient data, both in terms of

sample size and richness (such as variety of modalities and time points with behavioral measurements available for each patient). Modern machine learning methods show great promise in predictive capabilities but often require tens of thousands of samples for training. Adolfs, in his discussion about the future of lesions studies [3] states that for more accurate modeling and predictions, larger cohorts as well as more complex multivariate modeling approaches (for e.g., considering white and grey matter distinctly) are needed. In light of data scarcity, recent studies have made attempts to generate synthetic data (data augmentation) in biomedical databases to improve diagnostic performance [30] [50]. However, to the best of our knowledge, data augmentation methods have not been applied in the context of stroke yet. In conclusion, the development of modern methodologies (multivariate modeling, non-linear modeling, neural networks and data augmentation) could potentially help overcome some of the limitations of the early LSM studies and open pathways for innovation in precision medicine tools.

3

Methodology

3.1 Participant Sample

1401 participants included in this study are patients diagnosed with acute ischemic stroke in two South Korean hospitals, Hallym University Sacred Heart Hospital and Seoul National University Bundang Hospital, between January 2007 and December 2018 [36]. The mean (± one standard deviation) age of the participant sample is 67.7 (± 11.6) and 58% of them are male. The infarction is observed on the diffusion-weighted magnetic resonance imaging (DW-MRI) and fluid attenuated inversion recovery (FLAIR) sequences of the admitted patients carried out within 7 days in most cases of experiencing symptoms (The Bundang cohort scanned patients within 48 hours, while the Hallym cohort scanned around 7 days post-stroke). Patients were chosen in this study based on the the following criteria: (1) existence of visible acute infarct(s) on the diffusion-weighted imaging (DWI) or FLAIR, (2) unseen prior cortical infarcts, subcortical infarct(s) larger than 15 mm or hemorrhages larger than 10 mm, (3) successful registration and segmentation of the infarct, and (4) availability of the cognitive assessment scores (the 60-min Korean-Vascular Cognitive Impairment Harmonization Standards-Neuropsychology Protocol, K-VCIHS-NP; [28] [64]) and clinical data within a year of stroke onset. Patients who (1) had bilateral stroke, (2) due to severe aphasia could not take the neuropsychological assessment, and (3)

have inadequate MRIs (insufficient to obtain neuroimaging variables of interest) were excluded from this study. Data was acquired prospectively following the approved study protocols by the Institutional Review Boards of both hospitals.

3.2 Neuropsychological assessment

After 3 months of the onset of stroke, patients went through a series of behavioral assessments: the K-VCIHS-NP (median time post-stroke: 98 days; [64]). Performance in the following cognitive domains was used in this study:

1. Global cognition as measured by Korean version of the Mini-Mental State Examination (MMSE) (Total score; [26]). MMSE captures general cognitive performance by assessing a wide variety of domains such as registration, orientation to time and place, attention, calculation and language.

2. Language performance as measured by Korean short version of the Boston Naming Test (BN) (Total score; [35]). This is a standardized test to assess a patient's ability to name various objects.

3. Memory function as measured by the Seoul Verbal Learning Test (SVL), the Korean equivalent to the Hopkins Verbal Learning Test (Immediate recall; [10]). This test assesses especially the short term memory and learning ability of patients.

4. Visuospatial Functioning as measured by the Rey–Osterrieth Complex Figure Test (Copy; [47] [41]). This test assesses a patient's ability to recognize and reproduce a complicated line drawing.

5. Information Processing Speed as measured by the Korean Elderly version of the Trail Making Test (Part A; [42]). This test assesses a patient's visual search speed by drawing lines to connect circles in a numerical sequence (1, 2, 3, ...).

6. Executive Functioning as measured by the Korean Elderly version of the Trail Making Test (TMT) (Part B; [42]). This test assesses a patient's visual search speed and mental flexibility by drawing lines to connect circles in a sequence alternating between numbers and letters (1, A, 2, B, ...).

Table 3.1 describes the summary of the scores participants achieved in all the six neuropsychological assessments. Table 3.2 shows the more in-depth variation of the scores of all patients.

Table 3.1: Overview of the neuropsychological assessment scores. Summary statistics includes the count, number of missing values, mean, median, standard deviation and the range of the six cognitive assessments scores.

	Korean MMSE (Total score)	Korean Elderly TMT A	Korean Elderly TMT B	Korean BN (Total score)	Rey Complex Figure Test (Copy)	SVL (Immediate Recall)
N-valid	1397.00	1194.00	1076.00	1372.00	1289.00	1381.00
N-Missing	4.00	207.00	325.00	29.00	112.00	20.00
Mean	23.74	40.74	88.44	9.93	26.03	14.89
Median	26.00	30.00	56.50	10.00	29.00	15.00
Std. Dev	5.93	38.47	80.47	3.61	8.99	6.03
Range	30.00	294.00	289.00	15.00	36.00	36.00
Minimum	0.00	6.00	11.00	0.00	0.00	0.00
Maximum	30.00	300.00	300.00	15.00	36.00	36.00

Table 3.2: Distribution of the neuropsychological assessment scores. 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles of the six cognitive assessments scores.

Percentiles	5	10	25	50	75	90	95
Korean MMSE (Total score)	19.00	21.00	25.00	27.00	28.00	29.00	30.00
Korean TMT A	12.00	14.00	19.25	27.00	40.00	59.00	80.00
Korean TMT B	20.00	24.00	35.00	55.00	103.75	216.50	300.00
Korean BN (Total score)	6.00	7.00	9.00	12.00	13.00	14.00	15.00
Rey Complex Figure Test (Copy)	13.83	19.00	26.00	31.00	33.00	35.00	35.00
SVL (Immediate Recall)	8.00	9.00	13.000	16.00	20.00	24.00	25.35

3.3 Neuroimaging data

Structural axial T1, T2-weighted spin echo, DWI and FLAIR sequences (3.0T, Achieva scanner, Philips Healthcare, Netherlands, image dimensions: 182x218x182; see Table 3.3 and Table 3.4 for details) were included in the brain scanning. In house software based on MeVisLab (MeVis Medical Solutions AG, Bremen, Germany; [48]) were used to manually segment lesions in DWI (or FLAIR sometimes) by trained professionals (A.K.K., G.A.). Moreover, the segmentations were then examined and manually adjusted, where necessary, by two experienced raters (N.A.W., J.M.B). Using the RegLSM processing (public code: http://lsm.isi.uu.nl/; [58]), the segmented images were normalized, linearly and non-linearly, to the Montreal Neurological Institute (MNI-152) space. Following that, the registered lesion maps were examined for any visual discrepancies from the original image and manually corrected accordingly. Figure 3.1 shows a few sample processed images used in this analysis where the lesionized brain regions are highlighted.

3.4 Region of Interest based regression analyses

Infarct volumes (absolute size in mm^3) in a total of 193 brain regions were calculated based on the parcellations provided by following four widely used atlases:

1. Harvard Oxford Cortical Atlas (Threshold 50, Resolution 1mm, 93 regions) [22],
2. Harvard Oxford Subcortical Atlas (Threshold 50, Resolution 1mm, 18 regions) [22],
3. Cerebellum Atlas (34 regions) [24],

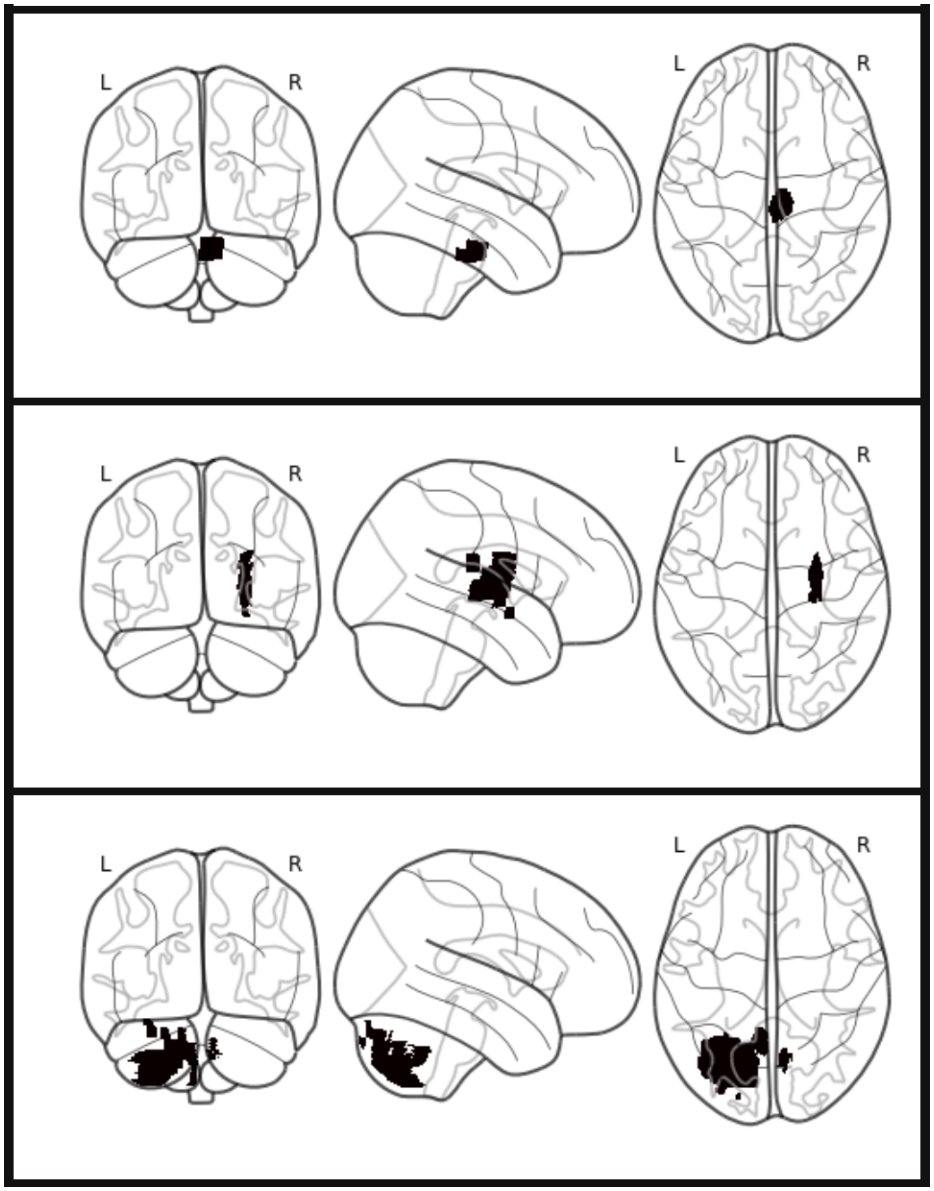

Figure 3.1: Coronal, sagittal and axial slices of three random samples of lesion segmented brain MRIs. The processed image data is binary where voxels with value=1 (colored black) indicate the presence of stroke lesion as identified by an expert.

Table 3.3: Details of the imaging protocol used in Seoul National University Bundang Hospital

FLAIR	• Repetition time: 11,000 ms; • Echo time: 125 ms; • Inversion time: 2,800 ms; • Slice thickness 5 mm; • Intersection gap: 1 mm; • Matrix: 512 x 512; • Flip angle 90 degree
DWI	• EPI-spin echo sequence; • Repetition time: 5,000 ms; • Echo time: 50 ms; • Diffusion b-value: 1,000; • Slice thickness: 5 mm; • Intersection gap: 1 mm; • Matrix: 256 x 256; • Flip angle 90 degree

4. International Consortium of Brain Mapping (ICBM) White-matter tractography atlas (48 regions) [37].

Figure 3.2 provides an overview of the prevalence of lesions in the dataset. We performed log transformation (base 10) of infarct volumes in these pre-defined brain regions in order to reduce the skewness in lesion volume distribution due to a larger number of small lesions. The log transformed values, z-scored, provided 193 input features to the regression based analyses. With this aggregation approach, the high dimensional voxel-wise data consisting about 1.8 million features is brought down to

Table 3.4: Details of the imaging protocol used in Hallym University Sacred Heart Hospital

FLAIR	• Repetition time: 11,000 ms; • Echo time: 125 ms; • Inversion time: 2,800 ms; • Slice thickness: 5 mm; • Matrix: 512 x 512; • Flip angle 90 degree
DWI	• Repetition time: 3,000 ms; • Echo time: 56 ms; • Diffusion b-value: 1,000; • Slice thickness: 5 mm; • Matrix: 256 x 256; • Flip angle 90 degree

a more manageable feature space of 193 dimensions, thereby mitigating the curse of dimensionality.

The output variables were the six different cognitive assessment scores, each z-scored. The missing data was filled with the simple random imputation method where each missing sample is replaced by randomly selecting one of the true (measured) scores of the respective cognitive domain.

After data pre-processing, we explored various machine learning models and techniques to predict cognitive scores from the lesion loads (lesion volumes in the aforementioned brain regions) systematically in three different dimensions. First, we compared single versus multi-outcome models. Second, linear versus non-linear models. Lastly, we apply the Mixup technique on the dataset. The analyses in this section were

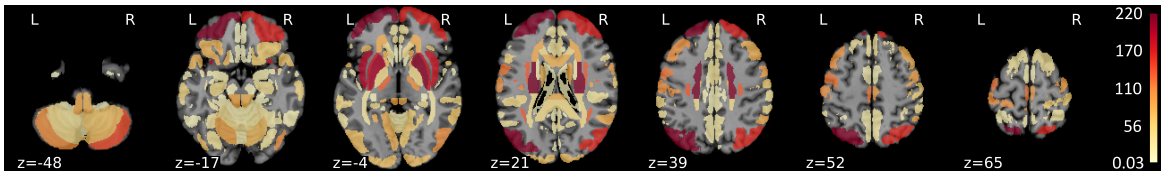

Figure 3.2: Lesion prevalence map. The color scale indicates the average lesion size (absolute volume in mm^3) in each brain region in 1401 patients in the dataset. All the 193 atlas derived regions are colored in this figure, and in each of these regions, at least one patient has a lesion. Brain images are projected on the 1 mm MNI-152 template (Z coordinates: -48, -17, -4, 21, 39, 52, 65). Lesion maps are displayed in the neurological convention (left brain is on the left (L) and right brain is one the right (R)).

performed in the Python 3.7 environment relying predominantly on the packages Nilearn (version 0.5.2) [2] and Scikit-learn (version 0.22.1) [44]. The code is available at `https://github.com/hasnainmamdani/stroke-impairment-analysis`.

3.4.1 Single vs Multi-Outcome modeling

The single-outcome models are those that model one cognitive outcome at a time. Hence, six models are needed to predict six different cognitive functions. The models used in this part of the regression analysis are Ridge (a particular type of Tikhonov regularization where the regularization is given by the l2-norm) [32], Support Vector Machines with Radial Basis Function kernel (SVR-RBF) [16], and Random Forest [11] [27].

Multi-outcome models train on and predict all the six cognitive functions with a single model. The multi-outcome models used here are Multitask Ridge, Partial Least Squares (PLS) [59] [55], Canonical Correlation Analysis (CCA) [59] [55] and (Multi-output) Random Forest [11] [27]. Note we used Random Forest for both single-outcome and multi-outcome analyses as Random Forest natively supports modeling of single as well as multiple outputs.

3.4.2 Linear vs Non-Linear modeling

The linear model used for comparison here is Ridge. In Ridge, the hyperparameter tuned was the i) regularization strength. The non-linear models used in this analysis are Support Vector Machines with Radial Basis Function kernel (SVR-RBF), and Random Forest. These are widely used models and suitable for benchmarking predictive performance [31].

3.4.3 Mixup

With mixup, we upsized the data to have a sample size of i) 5x and ii) 10x the original dataset size. For instance if the size of training data is 100 samples, a multiplication factor of 5x will generate 400 additional samples rendering a total sample size of 500. The interpolation parameter lambda used to construct additional samples was sampled from a Beta distribution with shape parameters (alpha, alpha). Alpha controls the interpolation strength between two randomly chosen samples to generate a new sample. For each of these settings, we tried various values of alpha: 0.01, 0.1, 0.3 and 1.0. Data augmentation was performed on the lesion volumes in ROIs before taking their logarithm. The models used with post Mixup data were Ridge and Random Forest. Note, data augmentation was only applied on the training data and prediction performance was measured on the unseen true samples.

In all the analyses, for each model type, nested cross validation scheme was utilized to obtain training (in-sample) and testing (out-of-sample) results. 1401 samples were divided into 5-folds. The outer loop was iterated 5 times, each time with a different data fold as test data. Within each iteration of the outer loop, 5-fold inner cross validation was further performed on the training data for hyperparameter tuning. The in-sample and out-of-sample estimates of coefficient of determination (r-square) were calculated and reported for each iteration of the outer loop.

In Ridge, the hyperparameter tuned was the i) regularization strength. In SVR-RBF, the hyperparameters tuned were the i) regularization strength and ii) epsilon, the width of the allowable error where samples predicted within that error range are not penalized during model training. In Random Forest, the hyperparameters tuned were the i) number of trees in the forest, ii) maximum number of features taken into account to determine the best split at a node, iii) maximum height of the trees, iv) minimum number of samples needed to split a node, v) minimum number of samples needed at the leaf node, and vi) maximum number of samples used to form a tree with bootstrapping. In Multitask Ridge, the hyperparameter tuned was the 1) regularization strength. In PLS, the hyperparameter tuned is the 1) number of components kept in the decomposition. In CCA, the hyperparameter tuned is the 1) number of components kept in the decomposition.

4

Results

We used machine learning models to predict cognitive functions three months post-stroke from the lesion topography. Here, we present results of the three dimensions of this study: single-outcome vs multi-outcome modeling; linear vs. non-linear modeling; and data augmentation using Mixup.

4.1 Single-Outcome vs. Multi-Outcome modeling

First, we compared the performance of multi-outcome models with single-outcome models. Figure 4.1 shows the out-of-sample coefficient of determination (r-square) values of various single and multi-outcome models. In general, for both types of modeling, we see the r-square scores are low indicating poor predictive performance and the error bars, which represent the standard deviation of r-square values across k-folds, are large showing lack of consistency in predictions. It is also noticeable that information processing speed and executive functioning domains are hard to predict from lesion topography. Having said that, at the sample size that we have available, we see subtle indicators that multi-output models provide slightly better mean predictive performance compared to single-output models, mostly in the domains

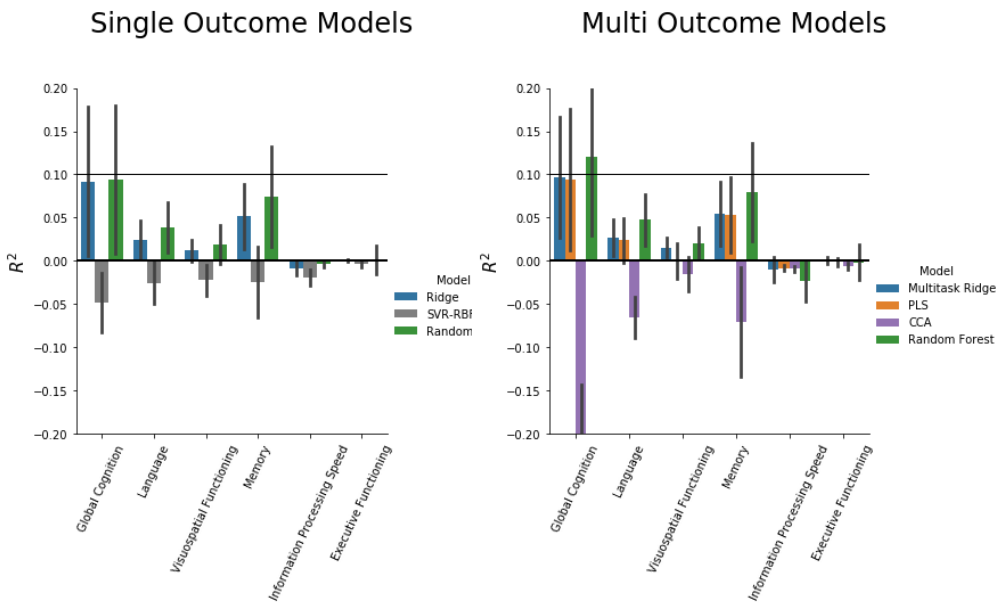

Figure 4.1: Mean out-of-sample coefficient of determination (r-square) values of various single and multi-output models. The error bars indicate the standard deviation of the r-square values across 5 folds for each model and cognitive domain. A line at $R^2 = 0.10$ is drawn for the ease of visualization. The mean predictive accuracy of the best performing multi-outcome model (Random Forest) is slightly better than that of the best performing single-outcome model (Random Forest) in the domains of global cognition and language.

of global cognition and language. The mean (± one standard deviation across cross-validation folds) r-square of the best performing multi-outcome model, Random Forest, in the domains of global cognition and language are 0.120 (± 0.101) and 0.047 (± 0.033) respectively. Whereas for the best performing single-outcome model, also Random Forest, the mean (± one standard deviation across cross-validation folds) r-square for global cognition and language domains are 0.094 (± 0.096) and 0.039 (± 0.033) respectively. This suggests that multi-output models may have capability to beneficially capture shared patterns across various clinical dimensions in a predictively useful fashion. Therefore, joint modeling of interrelated cognitive functions exhibit potential to perform more accurate single subject predictions. In-sample coefficient of determination (r-square) values of these single and multi-outcome models can be found in Appendix A.1.

4.2 Linear vs. Non-linear modeling

Secondly, we investigated whether non-linear models predict cognitive functions post stroke better than linear models. Fig 4.2 compares the out-of-sample coefficient of determination (r-square) values for various linear and non-linear models. It can be seen that despite low and inconsistent predictions in general, the mean predictive accuracy of the best performing non-linear model is slightly better compared to the best performing linear model for the Language and Memory functions. The mean (± one standard deviation across cross-validation folds) r-square of the best performing non-linear model, Random Forest, in the domains of language and memory are 0.039 (± 0.033) and 0.074 (± 0.065) respectively. Whereas for the linear model Ridge, the mean (± one standard deviation across cross-validation folds) r-square for language and memory are 0.025 (± 0.024) and 0.052 (± 0.042) respectively. The results hint that the performance of language and memory functions in stroke patients might be modeled more appropriately by considering a non-linear interaction between spatially distributed brain regions. Hence, there seems to be some evidence that non-linear modeling leads to slightly better single subject predictions on average. In-sample coefficient of determination (r-square) values of these linear and non-linear models can be found in Appendix A.2.

4.3 Mixup

Next we applied Mixup (cf. Methods), a data augmentation technique, to examine whether additional data generated by linear interpolation of true samples can help improve predictive performance of various machine learning models. Fig 4.3 shows the out-of-sample coefficient of determination (r-square) values for Ridge and Random Forest regression with different configurations of data augmentation. Both models, trained on post mixup augmented data, do not seem to perform better when tested

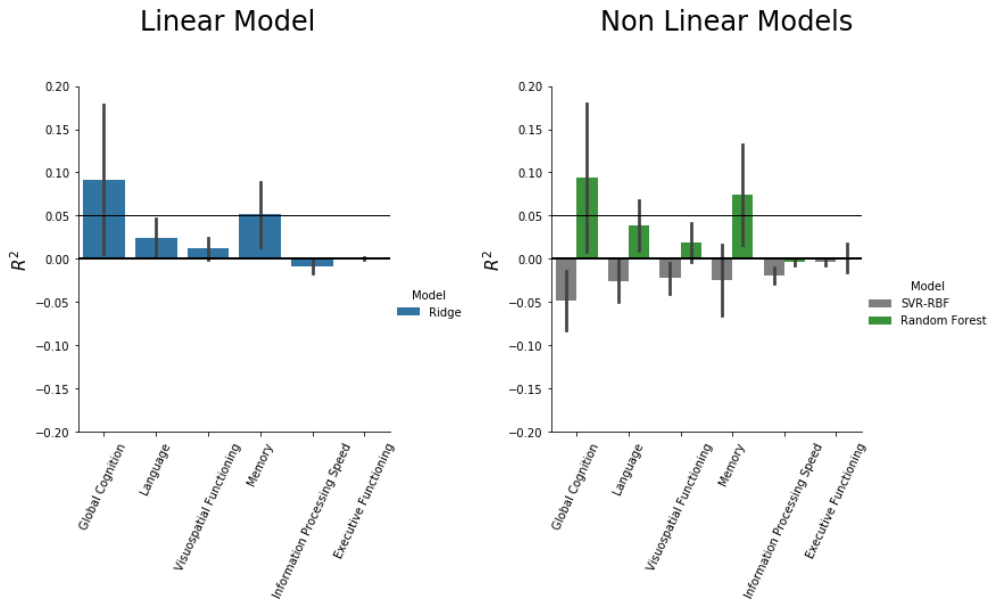

Figure 4.2: Mean out-of-sample coefficient of determination (r-square) values of various single-output linear and non-linear models. The error bars indicate the standard deviation of the r-square values across 5 folds for each model and cognitive domain. A line at $R^2 = 0.05$ is drawn for the ease of visualization. The predictive accuracy of the best performing non-linear model is slightly higher compared to the linear model in the domains of language and memory.

on completely unseen data compared to when they were trained on original training data only. For all the six domains, the mean r-square is lower when the models are trained on augmented training data compared to when the models are trained on original training data. Moreover, we notice a decline in predictive performance with the increase of data multiplication factor. With increasing values of alpha (the newly generated samples are more concentrated in the middle of the linear interpolation line between the two source samples, i.e., further from any one of the true samples), the predictive performance improves but it still is not close to the performance of the model trained on original training data only. Higher values of multiplication factor and alpha were also tried but the results were not very different than the ones reported. From these results, we see that it is not easy to exploit artificial samples generated by Mixup to improve predictive performance of cognitive functions post stroke. Therefore, with this stroke dataset, we are yet to make beneficial use of the

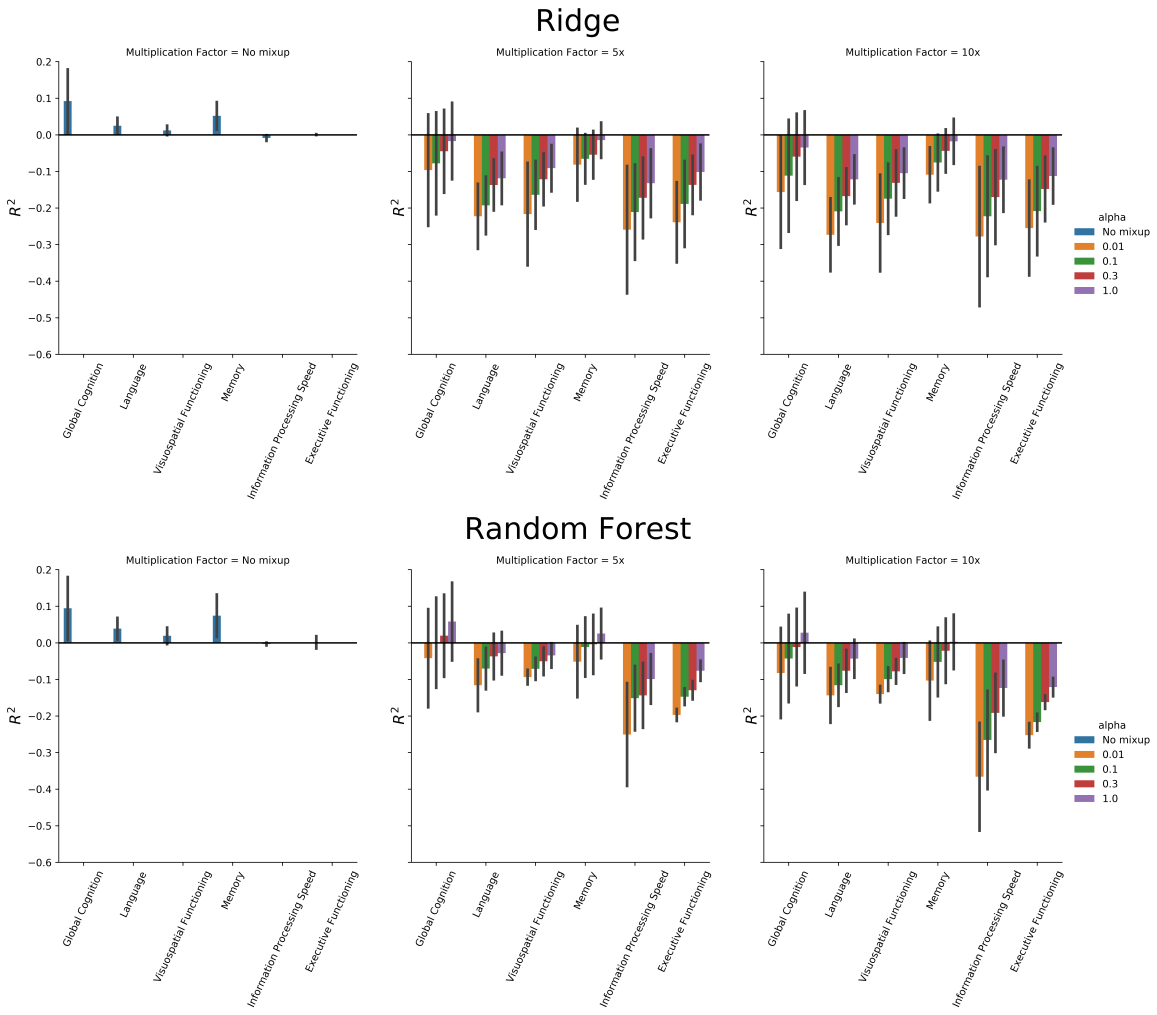

Figure 4.3: Mean out-of-sample coefficient of determination (r-square) values of the Ridge and Random forest regression model. In each figure, the left panel shows results without mixup i.e., no data augmentation, the middle panel shows results with mixup with 5x data augmentation and the right panel shows results with mixup with 10x data augmentation. The error bars indicate the standard deviation of the r-square values across 5 folds for each model and cognitive domain.

power of data augmentation in a predictive modeling setting. In-sample coefficient of determination (r-square) values of these models can be found in Appendix A.3.

5

Discussion and Future Work

We performed stroke outcome prediction analysis on a multisite database of 1401 stroke patients. A core strength of this study is the relatively large and diverse sample size. Previous studies on stroke outcome prediction are limited to sample sizes of at most a few hundred subjects [17] [38] [1] [67] [39] . Moreover, the neuropsychological assessments cover a wide variety of cognitive domains including measures for global cognition, language, memory, visuospatial functioning, information processing speed and executive functioning. High coverage of neurological lesions including regions of both gray and white matter and cognitive assessments allow us to, at the first place, reliably employ multi-outcome modeling, non-linear modeling and data augmentation in this study.

Firstly, we explored the utility of multi-outcome modeling in cognitive outcome prediction. The results suggest tentative hints that multi-output models might be able to better capture similar underlying disease processes of many clinical dimensions typically observed in stroke patients from a predictive modelling perspective.

Secondly, we compared the results of non-linear models with linear models. Non-linear models performed slightly better than the linear models in more than one cognitive domain in the prediction regime as shown by the mean variance explained (r-square) values. Hence, there seems some evidence that non-linear models are able

to potentially exploit some part of non-linearity that is found in cognitive processes leading to slightly better single subject predictions on average.

Thirdly, we analyzed whether augmenting data with the Mixup technique helps with improving cognitive outcome prediction of unseen samples. Even though we had a large and diverse bank of data to begin with, the predictive performance of various models decline when trained on augmented data suggesting that the relationship between lesion volumes and behavioral scores is not easily replicated by a simple linear interpolation method that Mixup employs. To our knowledge, this is the first study that explores the application of a modern data augmentation method in the context of stroke.

Many studies on stroke lesion symptom mapping focus on inference, i.e., providing mechanistic insights into brain regions that are associated with various cognitive functions [18] [62] [25] [1]. Inference based studies are, however, not necessarily centered on clinical endpoint prediction from a precision medicine perspective [43] [12]. In this study, we focused on predictions of cognitive outcomes post-stroke, which could assist clinicians in devising an individualized treatment strategy for each new patient. We measured the out-of-sample predictive performance of models by testing them on previously unseen subjects. On that note, we noticed that the stroke literature is inconsistent with the use of term prediction and suggest that 'predictions' should involve evaluation of models on data that is previously not used for the training/fitting or hyperparameter tuning of the model. Maintaining a focus on predictions allowed us to examine a multitude of modeling options, which is not the case in inference based studies where investigators are limited to employing models that have interpretable parameters only. We thoroughly tried various models from simple linear to complex non-linear and multi-output models to exploit the predictive information present in brain images. Another aspect of this study is that the prediction of cognitive outcomes post-stroke is based entirely on neuroimaging features and the influence of

socio-demographic factors such as age, gender, education, etc. is not modeled. Note, inspired from other related studies we tried using principal components as features [29] [38] [54] instead of lesion volumes in brain regions as well as artificial neural network models [17], but the results obtained were not much different than the ones reported. The reason behind using atlas based features as inputs to the regression models is that atlases are by design created based on the association of its ROIs with brain functioning.

Overall, the results are suggestive of improvement. We see that cognitive outcome prediction from lesion topography is a very challenging real-world problem from a predictive modelling perspective. Throughout the study, we notice the prediction accuracies as measured by coefficient of determination (r-square) are generally quite low and inconsistent across cross validation folds as reflected by large standard deviation of r-square values. This is in line with previous studies that suggest that the classic lesion locations explain chronic behavioral outcomes but with little explained variance [46]. There can be various explanations to that.

The notion of basing brain-behavior association on coarse spatially distributed patterns of lesions, which is the fundamental assumption made in most MRI based lesion symptom mapping studies, is arguable. Any spatial region in the brain may not be necessarily associated with a particular task [6]. Neurons are a general purpose computation unit and individuals may perform a certain cognitive function using different topographical regions of the brain [8] [53]. Hence, by standardizing brain region volumes to a common space, we might not be able to capture interindividual variations.

Moreover, MRIs are not able to precisely capture the infarcted regions of the brain's nervous system [40]. With MRIs, our lowest unit of measurement is a 1 mm^3 (1 mm x 1 mm x 1 mm) voxel. With this resolution, we can not peek into the infarction state

of individual neurons and axons (with diameters in tens of micrometers) and study their contribution to cognitive functioning. Hence, it is possible that the anatomical details not captured by MRIs could be a key missing ingredient required to predict cognitive impairments more accurately.

According to common criticism, MRIs in general contain limited information that can be exploited for predicting disease phenotypes. Imperfect brain imaging measurements need to be complemented by larger amounts of data to fully exploit the predictively useful information in them [61] [52]. Therefore, with the current size of the dataset, even the simple linear models may not have reached plateaus of performance. This is in accordance with the view that limited sample size is the bottleneck in image based predictions of neurological disorders [4].

Another reason for poor predictive performance could be that the cognitive outcomes measured post-stroke provide little indication of a person's cognitive abilities prior to stroke. That is to say, we have an absolute measure of cognitive performance post-stroke, but we do not have information about the decline in cognitive performance due to stroke. Without this, the models are not able to distinguish at an individual level the component of cognitive impairment caused by stroke and general cognitive capacity. Moreover, our data repository consists of patients suffering from a stroke only and as such there are no healthy controls. Lack of knowledge of how healthy participants perform on cognitive assessments would limit and partly skew the distinctive power of models. In other words, impairments can only be characterized so much without the knowledge of a desirable (healthy) state.

With the current quality of brain imaging measurements (which warrants extensive research in their progress), we see that a sample size of about fourteen hundred is not large enough to exploit predictively useful information in them for usefulness in a clinical setting. Given the severity and frequency of this neurological disorder, a

global effort needs to be coordinated to accomplish a data size of tens of thousands of stroke patients with consistent configurations of imaging protocols and measurements of cognitive abilities. The challenge in this century lies not in the scarcity of potential data but coordinating a consortium to collect and organize data from multitude of sites [3].

Besides increase in sample size, another direction to improving predictive power of models could be having more data per patient. One way to do so is by complementing structural MRIs with other modalities such as functional MRIs and electroencephalograms (EEG). Some studies have shown relevance of fMRIs to cognitive functions. For instance, global cognition reflects a combination of various cognitive functions, which depend on not only structural but also functional brain measurements [54]. Moreover, task based fMRI is shown to be useful in the prediction of language outcomes [51]. Future studies can explore contributions of multiple modalities towards prediction of cognitive outcomes.

Another way to increase single subject data is by collecting cognitive assessment scores of stroke patients over a period of time. The longitudinal data could help model and predict personalized trajectories of cognitive performance throughout the recovery phase and subsequently aid in treatments [33]. Lastly, binary lesion segmentations can be complemented with continuous imaging measurements that provide fine details of the severity of infarct. Continuous diffusion tensor imaging (DTI) parameter maps have been shown to more accurately predict long term cognitive outcome than binary lesion segmentation maps [38] [49].

6

Conclusions

We translated and benchmarked state of the art machine learning methods on a relatively large stroke dataset to predict cognitive outcomes from lesion topography. We showed that 1) multi-outcome models could possibly be exploited to improve predictions of global cognition and language outcomes; 2) non-linear models show promise in capturing non-linear interactions between distinct regions of the brain to predict language and memory functions more accurately; and 3) data augmentation with Mixup does not necessarily create useful samples from a predictive modeling viewpoint. In conclusion, we demonstrate that prediction of single patient outcomes from lesion topography is a difficult task. This work highlights the challenges and provides useful directions to future research in lesion-behavior mapping.

Bibliography

[1] Hugo P Aben et al. “Extent to Which Network Hubs Are Affected by Ischemic Stroke Predicts Cognitive Recovery”. In: *Stroke* 50.10 (2019), pp. 2768–2774.

[2] Alexandre Abraham et al. “Machine learning for neuroimaging with scikit-learn”. In: *Frontiers in neuroinformatics* 8 (2014), p. 14.

[3] Ralph Adolphs. “Human lesion studies in the 21st century”. In: *Neuron* 90.6 (2016), pp. 1151–1153.

[4] Mohammad R Arbabshirani et al. “Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls”. In: *Neuroimage* 145 (2017), pp. 137–165.

[5] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. “Convex multi-task feature learning”. In: *Machine learning* 73.3 (2008), pp. 243–272.

[6] Dmitriy Aronov, Rhino Nevers, and David W Tank. “Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit”. In: *Nature* 543.7647 (2017), pp. 719–722.

[7] Elizabeth Bates et al. “Voxel-based lesion–symptom mapping”. In: *Nature neuroscience* 6.5 (2003), pp. 448–450.

[8] Marina Bedny et al. “Language processing in the occipital cortex of congenitally blind adults”. In: *Proceedings of the National Academy of Sciences* 108.11 (2011), pp. 4429–4434.

[9] Anna K Bonkhoff et al. "Generative lesion pattern decomposition of cognitive impairment after stroke". In: *Available at SSRN 3682001* (2020).

[10] Jason Brandt. "The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms". In: *The Clinical Neuropsychologist* 5.2 (1991), pp. 125–142.

[11] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[12] Danilo Bzdok and John PA Ioannidis. "Exploration, inference, and prediction in neuroscience and biomedicine". In: *Trends in neurosciences* 42.4 (2019), pp. 251–262.

[13] Danilo Bzdok et al. "Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015, pp. 3348–3356.

[14] Statistics Canada. *Table 13-10-0394-01. Leading causes of death, total population, by age group.* URL: `https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039401` (visited on 12/12/2020).

[15] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[16] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27.

[17] Sucheta Chauhan et al. "A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images". In: *Frontiers in neuroinformatics* 13 (2019).

[18] Bastian Cheng et al. "Influence of stroke infarct location on functional outcome measured by the modified rankin scale". In: *Stroke* 45.6 (2014), pp. 1695–1702.

[19] Elisheva R Coleman et al. “Early rehabilitation after stroke: a narrative review”. In: *Current atherosclerosis reports* 19.12 (2017), p. 59.

[20] Maurizio Corbetta, Joshua S Siegel, and Gordon L Shulman. “On the low dimensionality of behavioral deficits and alterations of brain network connectivity after focal injury”. In: *Cortex* 107 (2018), pp. 229–237.

[21] Maurizio Corbetta et al. “Common behavioral clusters and subcortical anatomy in stroke”. In: *Neuron* 85.5 (2015), pp. 927–941.

[22] Rahul S Desikan et al. “An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest”. In: *Neuroimage* 31.3 (2006), pp. 968–980.

[23] Mandip S Dhamoon et al. “Long-term functional recovery after first ischemic stroke: the Northern Manhattan Study”. In: *Stroke* 40.8 (2009), pp. 2805–2811.

[24] Jörn Diedrichsen et al. “A probabilistic MR atlas of the human cerebellum”. In: *Neuroimage* 46.1 (2009), pp. 39–46.

[25] M Ernst et al. “Impact of ischemic lesion location on the mRS score in patients with ischemic stroke: a voxel-based approach”. In: *American Journal of Neuroradiology* 39.11 (2018), pp. 1989–1994.

[26] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. ““Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician”. In: *Journal of psychiatric research* 12.3 (1975), pp. 189–198.

[27] Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. In: *Machine learning* 63.1 (2006), pp. 3–42.

[28] Vladimir Hachinski et al. “National Institute of Neurological Disorders and Stroke–Canadian stroke network vascular cognitive impairment harmonization standards”. In: *Stroke* 37.9 (2006), pp. 2220–2241.

[29] Ajay D Halai, Anna M Woollams, and Matthew A Lambon Ralph. "Predicting the pattern and severity of chronic post-stroke language deficits from functionally-partitioned structural lesions". In: *NeuroImage: Clinical* 19 (2018), pp. 1–13.

[30] Changhee Han et al. "Learning more with less: Conditional PGGAN-based data augmentation for brain metastases detection using highly-rough annotation on MR images". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 119–127.

[31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[32] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[33] Thomas MH Hope et al. "Predicting outcome and recovery after stroke with lesions extracted from MRI images". In: *NeuroImage: clinical* 2 (2013), pp. 424–433.

[34] H Jokinen et al. "Post-stroke cognitive impairment is common even after successful clinical recovery". In: *European Journal of Neurology* 22.9 (2015), pp. 1288–1294.

[35] E Kaplan, H Goodglass, and S Weintraub. "The Boston Naming Test. Lea & Febiger". In: *Philadelphia, PA* (1983).

[36] Beom Joon Kim et al. "Case characteristics, hyperacute treatment, and outcome information from the clinical research center for stroke-fifth division registry in South Korea". In: *Journal of stroke* 17.1 (2015), p. 38.

[37] Susumu Mori et al. *MRI atlas of human white matter*. Elsevier, 2005.

[38] Eric Moulton et al. “Multivariate prediction of functional outcome using lesion topography characterized by acute diffusion tensor imaging”. In: *NeuroImage: Clinical* 23 (2019), p. 101821.

[39] Fanny Munsch et al. “Stroke location is an independent predictor of cognitive outcome”. In: *Stroke* 47.1 (2016), pp. 66–73.

[40] Lauren J O’Donnell and Carl-Fredrik Westin. “An introduction to diffusion tensor image analysis”. In: *Neurosurgery Clinics* 22.2 (2011), pp. 185–196.

[41] Paul Alexandre Osterrieth. “Le test de copie d’une figure complexe; contribution a l’etude de la perception et de la memoire.” In: *Archives de psychologie* (1944).

[42] John E Partington and Russell Graydon Leiter. “Partington’s Pathways Test.” In: *Psychological Service Center Journal* (1949).

[43] Martin P Paulus. “Pragmatism instead of mechanism: a call for impactful biological psychiatry”. In: *JAMA psychiatry* 72.7 (2015), pp. 631–632.

[44] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[45] Mehdi Rahim et al. “Joint prediction of multiple scores captures better individual traits from brain images”. In: *Neuroimage* 158 (2017), pp. 145–154.

[46] LE Ramsey et al. “Behavioural clusters and predictors of performance during recovery from stroke”. In: *Nature human behaviour* 1.3 (2017), pp. 1–10.

[47] André Rey. “L’examen psychologique dans les cas d’encéphalopathie traumatique.(Les problems.).” In: *Archives de psychologie* (1941).

[48] Felix Ritter et al. “Medical image analysis”. In: *IEEE pulse* 2.6 (2011), pp. 60–70.

[49] Jane M Rondina, Chang-hyun Park, and Nick S Ward. “Brain regions important for recovery after severe post-stroke upper limb paresis”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 88.9 (2017), pp. 737–743.

[50] Muhammad Sajjad et al. “Multi-grade brain tumor classification using deep CNN with extensive data augmentation”. In: *Journal of computational science* 30 (2019), pp. 174–182.

[51] Dorothee Saur et al. “Combining functional and anatomical connectivity reveals brain networks for auditory language comprehension”. In: *Neuroimage* 49.4 (2010), pp. 3187–3197.

[52] Marc-Andre Schulz et al. “Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets”. In: *Nature communications* 11.1 (2020), pp. 1–15.

[53] Jitendra Sharma, Alessandra Angelucci, and Mriganka Sur. “Induction of visual orientation modules in auditory cortex”. In: *Nature* 404.6780 (2000), pp. 841–847.

[54] Joshua Sarfaty Siegel et al. “Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke”. In: *Proceedings of the National Academy of Sciences* 113.30 (2016), E4367–E4376.

[55] Michel Tenenhaus. *La régression PLS: théorie et pratique*. Editions technip, 1998.

[56] Salim S Virani et al. “Heart disease and stroke statistics—2020 update: a report from the American Heart Association”. In: *Circulation* (2020), E139–E596.

[57] JM Wardlaw et al. “What is the best imaging strategy for acute stroke?” In: *Health Technology Assessment (Winchester, England)* 8.1 (2004), pp. iii–180.

[58] Nick A Weaver et al. “The Meta VCI Map consortium for meta-analyses on strategic lesion locations for vascular cognitive impairment using lesion-symptom mapping: Design and multicenter pilot study”. In: *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* 11.1 (2019), pp. 310–326.

[59] Jacob A Wegelin et al. “A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case”. In: *University of Washington, Tech. Rep* (2000).

[60] Daniel Wiesen et al. “Using machine learning-based lesion behavior mapping to identify anatomical networks of cognitive dysfunction: spatial neglect and attention”. In: *NeuroImage* 201 (2019), p. 116000.

[61] Choong-Wan Woo et al. “Building better biomarkers: brain models in translational neuroimaging”. In: *Nature neuroscience* 20.3 (2017), p. 365.

[62] Ona Wu et al. “Role of acute lesion topography in initial ischemic stroke severity and long-term functional outcomes”. In: *Stroke* 46.9 (2015), pp. 2438–2444.

[63] Grigori Yourganov et al. “Predicting aphasia type from brain damage measured with structural MRI”. In: *Cortex* 73 (2015), pp. 203–215.

[64] Kyung-Ho Yu et al. “Cognitive impairment evaluated with Vascular Cognitive Impairment Harmonization Standards in a multicenter prospective stroke cohort in Korea”. In: *Stroke* 44.3 (2013), pp. 786–788.

[65] Hongyi Zhang et al. “mixup: Beyond empirical risk minimization”. In: *arXiv preprint arXiv:1710.09412* (2017).

[66] Yongsheng Zhang et al. “Multivariate lesion-symptom mapping using support vector regression”. In: *Human brain mapping* 35.12 (2014), pp. 5861–5876.

[67] Lei Zhao et al. “Strategic infarct location for post-stroke cognitive impairment: A multivariate lesion-symptom mapping study”. In: *Journal of Cerebral Blood Flow & Metabolism* 38.8 (2018), pp. 1299–1311.

Appendices

A

In-Sample Results

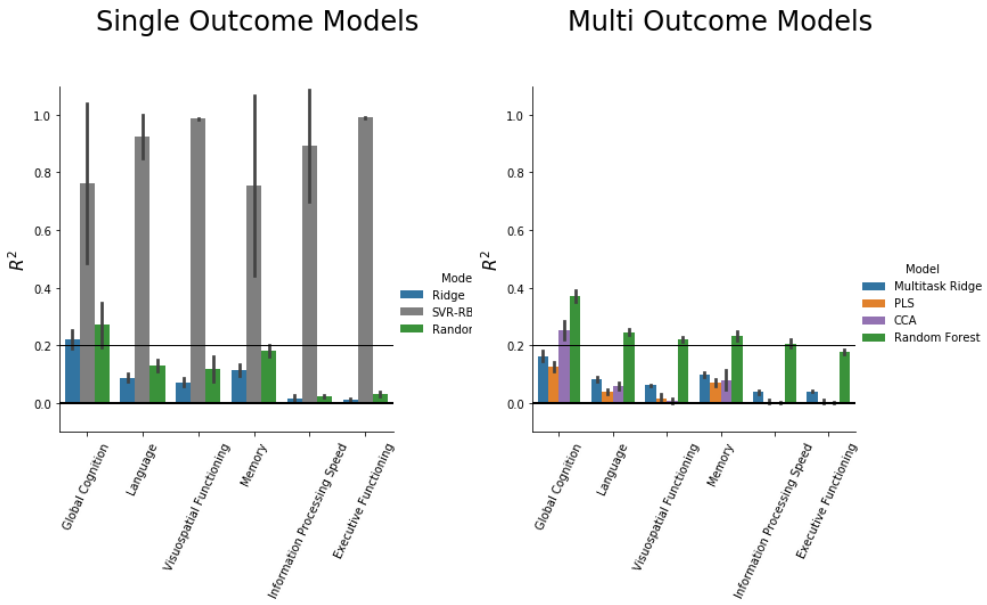

Figure A.1: Mean in-sample coefficient of determination (r-square) values of various single and multi-output models. The error bars indicate the standard deviation of the r-square values across 5 folds for each model and cognitive domain. A line at $R^2 = 0.20$ is drawn for the ease of visualization.

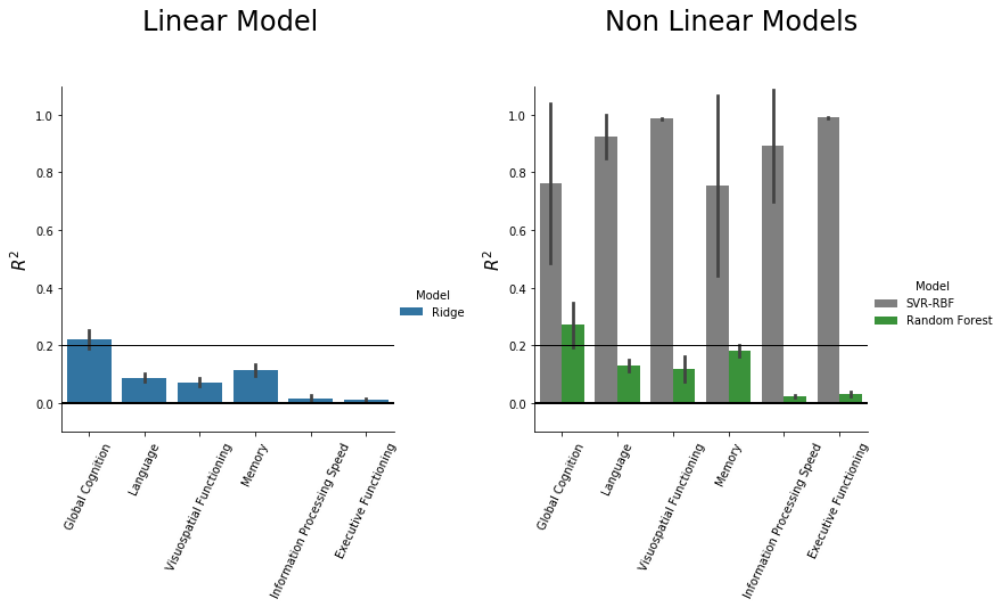

Figure A.2: Mean in-sample coefficient of determination (r-square) values of various single-output linear type and non-linear models. The error bars indicate the standard deviation of the r-square values across 5 folds for each model and cognitive domain. A line at $R^2 = 0.20$ is drawn for the ease of visualization.

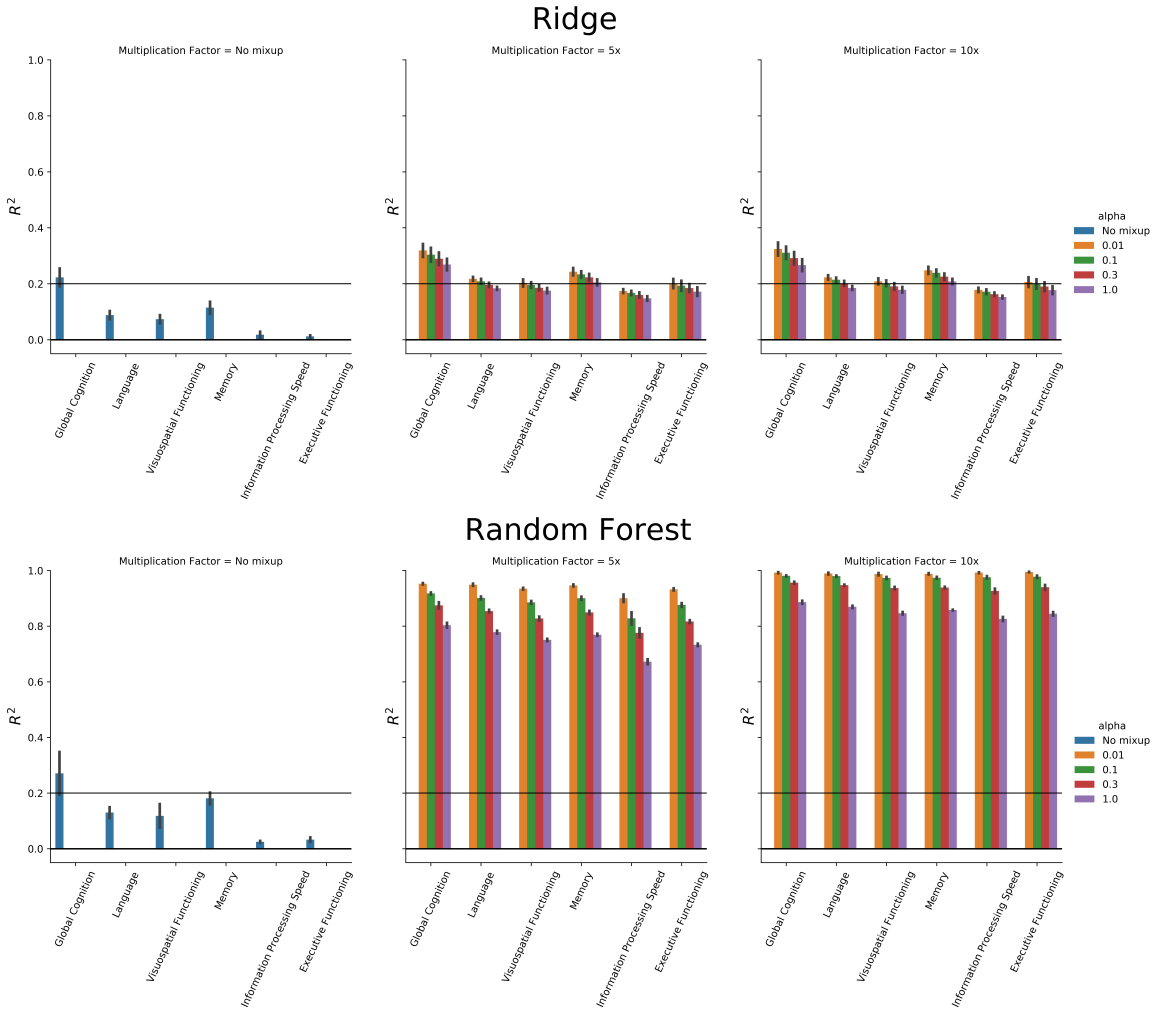

Figure A.3: Mean in-sample coefficient of determination (r-square) values of the Ridge and Random forest regression model. In each figure, the left panel shows results without mixup i.e., no data augmentation, the middle panel shows results with mixup with 5x data augmentation and the right panel shows results with mixup with 10x data augmentation. The error bars indicate the standard deviation of the r-square values across 5 folds for each model and cognitive domain. A line at $R^2 = 0.20$ is drawn for the ease of visualization.