Machine Learning and Deep Learning Applications in Neuroimaging

by
Gowtham Krishnan Murugesan

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy to the faculty of University of Texas at Arlington, University of Texas at Dallas and UT Southwestern Medical Center

Doctor of Philosophy
In
Biomedical Engineering

Supervising Committee:

Dr. Joseph Maldjian
Dr. Hanli Liu
Dr. Ananth Madhuranthakam
Dr. Sriraam Natarajan
Dr. Won Hwa Kim

Acknowledgment

It is my pleasure to acknowledge the roles of several individuals who were instrumental for the completion of my Ph.D. research.

First and foremost, I would like to express my sincere thanks to my supervising mentor Dr. Joseph Maldjian, who expertly guided me through my Ph.D. His unwavering enthusiasm in applying novel deep learning techniques to neuroimaging kept me constantly engaged with my research, and his personal generosity helped me make my time enjoyable at UT southwestern

I would like to extend my gratitude to Dr. Hanli Liu for supporting me right from my masters and whose personality and motivation inspired me to continue my research in neuroimaging. I also would like to thank Dr. Albert Montillo for giving me an opportunity to work in UT southwestern and mentoring me to develop machine learning skills.

I would like to thank the rest of my dissertation committee (Dr. Ananth Madhuranthakam, Dr. Won Hwa Kim, and Dr. Sriraam Natarajan) for their insightful comments and invaluable advice.

My appreciation also extends to my lab faculties Ben Wagner, Dr. Elizabeth Davenport, Dr. Fang F Yu, Dr. Thomas O Neil, Dr. Bhavya Shah, and colleagues Chandan Ganesh, Sahil Nalawade, James Holcomb and Divya Reddy for extending their help and support in all possible ways.

My gratitude goes to the faculty and staff of bioengineering departments at UT Arlington and UT Southwestern. Special thanks to Ms. Julia Rockow for all her help and encouragement from the day I arrived at UTA to the last day

August 11, 2020

Table of Contents

List of Figures

List of Tables

1. Overview of Deep Learning in Neuroimaging

1.1 Introduction

Over the past few decades, medical imaging techniques, such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), mammography, ultrasound, and X-ray, have been used for the early detection, diagnosis, and treatment of diseases. In the clinic, mostly human experts such as radiologists and physicians have performed medical image interpretation. However, given wide variations in pathology and the potential fatigue of human experts, researchers and doctors have begun to benefit from computer-assisted interventions. Although the rate of progress in computational medical image analysis has not been as rapid as that in medical imaging technologies, the situation is exponentially improving with the introduction of machine learning techniques. Machine learning is a technique for recognizing patterns that can be applied to medical images. Machine learning begins with computing the image features that are believed to be of importance in making the prediction or diagnosis of interest [1]. Recently deep learning applications have gained momentum as they do not require any manual image feature identification and calculation as a first step. Instead the features are identified as part of the learning process. The unprecedented success of deep learning is mostly due to the following factors: (a) advances in high-tech central processing units (CPUs) and graphics processing units

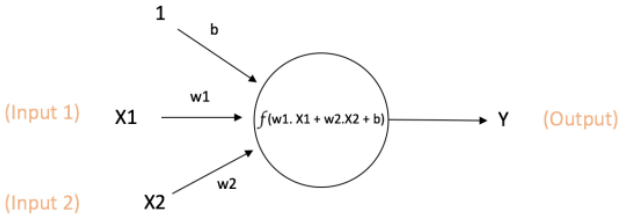


Figure 1: Representation of Single Artificial Neuron

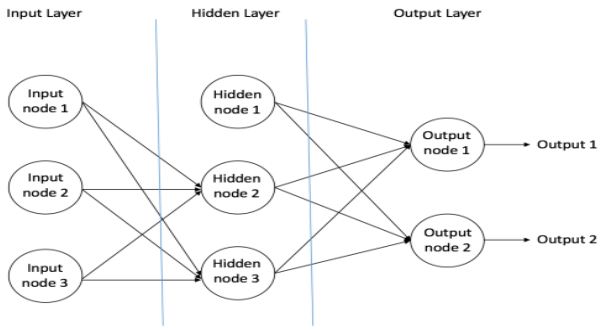


Figure 2: Artificial Neuronal Network (ANN) with one hidden layer

(GPUs), (b) the availability of a massive amount of data (i.e., big data), and (c) optimizations and developments in learning algorithms, such as U-net, auto-encoders, and generational adversarial networks for biomedical image processing [2].

Technically, deep learning can be regarded as an improvement over conventional artificial neural networks in that it enables the construction of networks with multiple (more than two) layers. An Artificial Neural Network (ANN) is a computational model that is inspired by the way biological neural networks in the human brain process information. The basic unit of computation in a neural network is the **neuron**, often called a node or unit. It receives input from some other nodes or an external source and computes an output. Each input has an associated weight (w), which is assigned based on its relative importance to other inputs. The node applies a function **f** (defined below) to the weighted sum of its inputs, as shown in Figures1 and 2. The network in Figure 1 takes numerical inputs **X1** and **X2** and has weights **w1** and **w2** associated with those inputs.

Additionally, there is another input **one** with weight **b** (called the **Bias**) associated with it. The feedforward neural network was the first and simplest type of artificial neural network devised. It contains multiple neurons (nodes) arranged in **layers**. Nodes from adjacent layers have **connections** or **edges** between them. All these connections have **weights** associated with them.

Figure 3: Representation of output from Sobel Filter

Convolutional Neural Networks (CNN) are the key resources in deep learning. They are based on a mathematical operation called convolution. A convolution is just a multiplication of an input image (which is a matrix) with a kernel (which is another matrix, aka filters) to generate a different image. In the example in Figure 3, a kernel called a Sobel filter has been applied over the Y-axis. The output image enhances the borders in the Y-axis (typically, in computer vision, the X-axis corresponds to the height and the Y-axis to the width). The key advantage of convolutions is that they produce new images, related to the original one, that enhance different areas of the input. They are featuarizers, or automatic feature generators. That is the key concept that makes deep learning so powerful.

A CNN is a neural network with many hidden layers that can use convolutions in any dimension (1D, 2D or 3D), to generate features hierarchically. The first layer or input layer convolves the input image to generate another set of images, usually called feature maps, creating a different representation of the input. The second layer, or first hidden layer, convolves the first feature maps to generate another set of feature maps, which are representations of the first hidden layer, and so on until the output layer is reached. Finally, a hierarchical structure is obtained where each layer contains images that are representations of the previous ones. Using several convolutional layers, deep neural networks can discover hierarchical feature representations such that higher-level features can be derived from lower-level features. Because these techniques enable hierarchical

feature representations to be learned solely from data, deep learning has achieved record-breaking performance in a variety of artificial intelligence applications and grand challenges (see https://grand-challenge.org).

1.2 Deep Learning applications in Neuroimaging

Improvements in computer vision prompted the use of deep learning in medical image analysis, such as image segmentation, lesion detection, image registration, image fusion, image annotation, computer-aided diagnosis (CADx) and prognosis, and lesion/landmark detection [3]. Possibly the most widespread application of machine learning-based diagnosis is in neurodegenerative diseases, where researchers aim to diagnose Alzheimer's disease or other forms of dementia or predict conversion from mild cognitive impairment (MCI) to dementia, based on brain MR images [4-6]. This is likely driven, at least in part, by the public availability of large datasets with diagnostic labels, such as the Alzheimer's disease Neuroimaging Initiative (ADNI), Brain Tumor Segmentation Challenge (BRATS) and Open Access Series of Imaging Studies (OASIS) [7-9]. Most initial deep learning applications in neuroradiology have focused on using computer vision techniques for detection and segmentation of anatomical structures and the detection of lesions, such as hemorrhage, stroke, metastases, aneurysms, primary brain tumors, and white matter hyperintensities. There are other innovative applications of AI in various technical aspects of medical imaging, including image acquisition, image registration, artifact correction, super-resolution, and dose reduction [10, 11].

1.3 Challenges in Applying Deep Learning Techniques in Neuroimaging

Deep learning has demonstrated excellent performance in automating several tasks in neuroimaging. However, there are some significant challenges and limitations. The significant challenges in applying the ML/DL techniques to the medical image analysis include:

1. Deep learning is an intensely data-hungry technology. It requires a vast number of well-labeled examples to achieve accurate classification and validate its performance for clinical implementation.
2. In the medical imaging domain, especially in neuroimaging, acquiring sufficient data for deep learning model development is a significant challenge. Collaboration between institutions could address this situation by sharing data. However, sharing medical data faces various legal, privacy, technical and data-ownership challenges. Several frameworks and techniques such as cyclic independent and identically distributed and federated learning between the institutions by passing the deep learning model between institution without sharing the data are being proposed to overcome such issues [12]
3. The neuroimaging data is mostly three dimensional (3D). Most of the models developed to date, and successful model architectures are in 2D space. High dimensional neuroimaging data is associated with memory and computational challenges. This is typically dealt with using smaller sized extracted patches or using 2D models on 3D data
4. Class imbalances and collecting a lot of data in smaller samples is another significant problem in medical imaging. Neuroimaging typically has wider datasets with more features collected on fewer samples than taller datasets, as well as inherentclass imbalances. For example, in brain tumor studies, there are vastly more normal subjects available than brain tumor subjects. Ideally, a dataset should have more numbers of samples (subjects) and fewer but important features (age, head impact exposure) collected on them, creating a

taller problem (more rows and fewer columns). Instead, in medical images such as MRI, more than a hundred thousand voxels data aree collected in a relative smaller number of subjects to identify a few hundred abnormal voxels (a wider problem, more columns than rows). This results in wider data sets with a large number of features collected on only a few subjects, causing the machine learning algorithms to overfit on the training data. In such cases, the application of ML algorithms requires feature selection and proper tuning of the hyperparameters of the model to avoid overfitting so that a model can generalize and behave the same as it did on training data. Training a complex classifier with a small dataset always carries the risk of overfitting. Deep learning models tend to fit the data exceptionally well, but this does not mean that they generalize well. Many studies used different strategies to reduce overfitting, including regularization, early stopping, and drop out). Overfitting can be evaluated by the performance of the algorithm on a separate test data set and by using cross-validation approaches. The algorithm may not perform well on similar images acquired in different centers, on different scanners, or with different patient demographics. Larger data sets from different centers are typically acquired in different ways using different scanners and protocols, with subtly different image features, leading to poor performance[13]. Overcoming this problem, known as “brittle AI,” is an essential area of research if these methods are to be used widely [10].

5. Deep networks are generally regarded as “black boxes” where data is input, and an output prediction, whether classification or image, is produced [13]. As deep learning algorithms rely on complicated interconnected hierarchical representation to produce its prediction, interpreting these predictions becomes very difficult, which has been coined “The Mythos

of Model Interpretability" [14]. Some estimates of the network uncertainly in prediction would be helpful to interpret better the images produced.

In summary, deep learning is a machine learning method that encompasses supervised, unsupervised, and semi-supervised learning. Deep learning methods demonstrate robust performance, often better than humans in various medical imaging applications. However, there are several challenges associated with the clinical translation of such algorithms. Despite the promises made by many studies, the reliable application of deep learning for neuroimaging remains in its infancy, and many challenges remain.

1.4 Dissertation Aims:

The goal of this dissertation is to apply advanced deep learning methods in three distinct domains of neuroimaging, 1. Exploratory Analyses, 2. Image Synthesis, and 3. Clinical applications in neuroimaging.

Exploratory Analyses

Machine learning models have been extensively applied in exploratory analyses to identify functional changes in the brain under various diseased conditions such as Attention Deficit/Heractivitiy Disorder (ADHD) and Alzheimer's Disease (AD). However, machine learning has not been used to infer functional networks in the brain. in this work, wewe developed a machine learning (ML) based inference methodfor functional MRI data (fMRI). called BrainNET. We validated the proposed model on ground truth simulation data. BrainNET outperformed standard correlation methods in terms of accuracy and sensitivity across simulations and various confounders such as the presence of cyclic connections, and even with truncated fMRI sessions of only 2.5 min. We evaluated the performance of BrainNET on the open-source "ADHD

200 preprocessed" data from the Neuro Bureau. BrainNET was able to identify significant changes in global graph metrics between ADHD groups and typically developing children (TDC), whereas correlation and partial correlation was unable to find any differences. The proposed method efficiently tackles the overfitting problem and considers each region of interest (ROI) as a separate feature and each time point as the sample, such that the resulting data has more columns than features.

Further, we applied BrainNET in an exploratory study to analyze the effects of head impact exposure (HIE) from subconcussive impacts on youth and high school (ages 9-18) football players to understand the functional network architecture in the human brain. In this study, we utilized graph theory,ML and data-driven methods to examine functional changes in the brain over a single season of American football. This study demonstrates an association between changes in functional connectivity related to HIE level in youth and high school football.

Image Synthesis

Second, we developed a novel deep learning algorithm to synthesize post gadolinium contrast images using only non-contrast MR images. In this study, we used novel deep learning approaches to synthesize T1 post-contrast (T1c) Gadolinium enhancement from non-contrast multi-parametric MR images (T1w, T2w, and FLAIR) in patients with primary brain tumors. Two expert neuroradiologists independently scored the synthesized post-contrast images using a 3-point scale (1, poor; 2, good; 3, excellent). The predicted T1c images demonstrated structural similarity, PSNR, and NMSE scores of 95.62 37.8357, and 0.0549, respectively. Our model was able to synthesize Gadolinium enhancement in 92.8% of the cases. Specific advantages of our strategy included the use of a more diverse dataset through the BRATS dataset. While introducing more

heterogeneity to the training dataset, this environment also enhances the generalizability of the trained networks. Further, standard preprocessing applied to all the data acquired with different clinical protocols and various scanners from multiple institutions make this approach useful for generalizing automated approaches across institutions, when differences in hardware and software can significantly alter image representations.

Clinical Applications

Finally, we developed DL algorithms to aid clinical applications of neuroimaging in identifying genetic mutation status in brain tumor patients and segmenting brain tumors with state-of-the-art results. For brain tumor segmentation, we developed multiple 2D and 3D segmentation models with multiresolution input to segment brain tumor components and then ensemble them to obtain robust segmentation maps. Ensembling reduced overfitting and resulted in a more generalized model. Multiparametric MR images of 335 subjects from the BRATS 2019 challenge were used for training the models. Further, we tested a classical machine learning algorithm with features extracted from the segmentation maps to classify subject survival range. Preliminary results on the BRATS 2019 validation dataset demonstrated excellent performance with DICE scores of 0.898, 0.784, 0.779 for the whole tumor (WT), tumor core (TC), and enhancing tumor (ET), respectively. The Ensemble of multiresolution 2D networks achieved 88.75%, 83.28%, and 79.34% dice for WT, TC, and ET, respectively, in a test dataset of 166 subjects. For brain tumor IDH mutation classification, we proposed an automated pipeline with minimal preprocessing and tested several state-of-the-art deep learning algorithms to predict IDH status noninvasively using deep learning and T2-weighted (T2w) MR images

The specific aims of the dissertation research are presented below:

1. Develop a machine learning based fMRI network inference method
2. Investigate the effect repetitive subconcussive head impact on youth and high school football players using data-driven and machine learning methods
3. Develop a deep learning method to synthesize post gadolinium contrast images with only using non-contrast images
4. Develop deep learning methods to achieve state-of-the-art brain tumor segmentation results
5. Apply state of the art deep learning algorithms to classify brain tumor IDH mutation status

1.5 Overview of Thesis

This document is an article-based dissertation with chapters 2-6 being referred to conference, journal and in preparation articles fulfilling the aims set out.

- Chapter 1: Overview of deep learning applications in neuroimaging is discussed along withthe challenges in applying such algorithms for clinical translation. Finally, a summary of contributions is provided with an overview of the dissertation.
- Chapter 2: A detailed description of a novel machine learning based fMRI network inference method called BrainNET is discussed based on the paper published in the Brain Connectivity Journal (2020)
- Chapter 3: A data-driven and machine learning based exploratory approach to identify changes in functional connectivity in the brains of young football players subjected to repetitive head impact exposure is discussed.
- Chapter 4. A novel deep learning approach to synthesize T1 post-contrast (T1c) Gadolinium enhancement from non-contrast multi-parametric MR images (T1w, T2w, and FLAIR) in patients with primary brain tumors is discussed in this chapter.

- Chapter 5: We developed ensemble deep learning algorithms of using multiple resolutions and multidimensional, deep learning algorithms in a clinical application to segment brain tumor into its sub-components with state of the art results.
- Chapter 6: We utilized several state of the art deep learning algorithms in a clinical application to predict IDH mutation status using T2w MRI alone.

1.6 References

[1] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics,* vol. 37, no. 2, pp. 505-515, 2017.

[2] M. Kim *et al.*, "Deep Learning in Medical Imaging," *Neurospine,* vol. 17, no. 2, pp. 471-472, Jun 2020, doi: 10.14245/ns.1938396.198.c1.

[3] J. Sui, M. Liu, J. H. Lee, J. Zhang, and V. Calhoun, "Deep learning methods and applications in neuroimaging," *J Neurosci Methods,* vol. 339, p. 108718, Jun 1 2020, doi: 10.1016/j.jneumeth.2020.108718.

[4] J. W. Song, N. R. Yoon, S. M. Jang, G. Y. Lee, and B. N. Kim, "Neuroimaging-Based Deep Learning in Autism Spectrum Disorder and Attention-Deficit/Hyperactivity Disorder," *Soa Chongsonyon Chongsin Uihak,* vol. 31, no. 3, pp. 97-104, Jul 1 2020, doi: 10.5765/jkacap.200021.

[5] A. Abrol *et al.*, "Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease," *J Neurosci Methods,* vol. 339, p. 108701, Jun 1 2020, doi: 10.1016/j.jneumeth.2020.108701.

[6] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong, "Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review," *Comput Methods Programs Biomed,* vol. 187, p. 105242, Apr 2020, doi: 10.1016/j.cmpb.2019.105242.

[7] C. R. Jack Jr *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine,* vol. 27, no. 4, pp. 685-691, 2008.

[8] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging,* vol. 34, no. 10, pp. 1993-2024, 2014.

[9] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience,* vol. 19, no. 9, pp. 1498-1507, 2007.

[10] G. Zhu, B. Jiang, L. Tong, Y. Xie, G. Zaharchuk, and M. Wintermark, "Applications of deep learning to neuro-imaging techniques," *Frontiers in Neurology,* vol. 10, p. 869, 2019.

[11] J. Jurek, M. Kociński, A. Materka, M. Elgalal, and A. Majos, "CNN-based superresolution reconstruction of 3D MR images using thick-slice scans," *Biocybernetics and Biomedical Engineering,* vol. 40, no. 1, pp. 111-125, 2020.

[12] M. J. Sheller *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific RepoRtS,* vol. 10, no. 1, pp. 1-12, 2020.

[13] F. Pesapane, C. Volonté, M. Codari, and F. Sardanelli, "Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States," *Insights into imaging,* vol. 9, no. 5, pp. 745-753, 2018.

[14] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31-57, 2018.

2 BrainNET: Inference of brain network topology using Machine Learning

Accepted: Brain Connectivity Journal

Authors: Gowtham Krishnan Murugesan, M.S[1], Chandan Ganesh, M.S[1], Sahil Nalawade, M.S[1], Elizabeth M Davenport, PhD[1], Ben Wagner, BM[1], Won Hwa Kim, PhD[2], Joseph A. Maldjian, MD[1]

Department of Radiology, UT Southwestern Medical Center[1]

Department of Computer science, The University of Texas at Arlington[2]

Author contributions: Gowtham Krishnan Murugesan conceptualized and designed the work, analyzed and interpreted the data, and wrote the paper. Dr. Joseph Maldjian and Dr. Won Hwa Kim provided expert knowledge and mentorship to develop the method. Ben Wagner contributed to develop fMRI analysis. Chandan Ganesh, Sahil Nalawade, and Dr. Elizabeth Davenport contributed to review the paper

The preprint version of this paper titled " BrainNET: Inference of brain network topology using Machine Learning" is available in 'biorxiv.org' (DOI: https://doi.org/10.1101/776641)

The paper is accepted in Brain Connectivity journal and reproduced in this chapter with permission.

2.1 Abstract

Background: To develop a new fMRI network inference method, BrainNET, that utilizes an efficient machine learning algorithm to quantify contributions of various regions of interests (ROIs) in the brain to a specific ROI.

Methods: BrainNET is based on Extremely Randomized Trees (ERT) to estimate network topology from fMRI data and modified to generate an adjacency matrix representing brain network topology, without reliance on arbitrary thresholds. Open source simulated fMRI data of fifty subjects in twenty-eight different simulations under various confounding conditions with known ground truth was used to validate the method. Performance was compared with correlation and partial correlation (PC). The real-world performance was then evaluated in a publicly available Attention-deficit/hyperactivity disorder (ADHD) dataset including 134 Typically Developing Children (mean age: 12.03, males: 83), 75 ADHD Inattentive (mean age: 11.46, males: 56) and 93 ADHD Combined (mean age: 11.86, males: 77) subjects. Network topologies in ADHD were inferred using BrainNET, correlation, and PC. Graph metrics were extracted to determine differences between the ADHD groups.

Results: BrainNET demonstrated excellent performance across all simulations and varying confounders in identifying true presence of connections. In the ADHD dataset, BrainNET was able to identify significant changes ($p < 0.05$) in graph metrics between groups. No significant changes in graph metrics between ADHD groups was identified using correlation and PC.

***Keywords*—Brain, Connectivity Analysis, fMRI, Machine Learning.**

2.2 Introduction

The brain is a complex interconnected network that balances segregation and specialization of function with strong integration between regions, resulting in complex and precisely coordinated dynamics across multiple spatiotemporal scales [15]. Connectomics and graph theory offer powerful tools for mapping, tracking, and predicting patterns of disease in brain disorders through modeling brain function as complex networks [16]. Studying brain network organization provides insight in understanding global network connectivity abnormalities in neurological and psychiatric disorders [17]. Several studies suggest that pathology accumulates in highly connected hub areas of the brain [18, 19] and that cognitive sequelae are closely related to the connection topology of the affected regions [20]. An understanding of network topology may allow prediction of expected levels of impairment, determination of recovery following an insult and selection of individually tailored interventions for maximizing therapeutic success [21]. A large number of network inference methods are being used to model brain network topology with varying degrees of validation. A recent study [22] evaluated some of the most common methods, including correlation, partial correlation, and Bayes NET, to infer network topology using simulated resting state functional magnetic resonance images (fMRI) data with known ground truth and found that performance can vary widely under different conditions.

Development of statistical techniques for valid inferences on disease-specific group differences in brain network topology is an active area of research [23, 24]. Machine learning methods have been used in neuroimaging for disease diagnosis and anatomic segmentation [25, 26]. Brain Network Construction and Classification (BrainNetClass) and GraphVar toolboxes provide a full pipeline from network construction to classification. BrainNetClass comprise various fMRI

network inference methods such as correlation, partial correlation, and higher-order functional connectivity for brain network inference followed by feature extraction for machine learning model development and testing [27, 28]. Very few studies have attempted to apply machine learning methods on direct time series of fMRI to infer brain networks [25, 29-31]. Recent work in machine learning approaches for inference of Gene Regulatory Networks (GRN) has demonstrated excellent performance [32-34]. Interestingly, these same approaches to gene regulatory networks can be used to infer brain networks. In this study, we describe a new network inference method called BrainNET, inspired by machine learning methods used to infer GRN [35].

Yan et al. devised a *bidirectional* Long Short-Term Memory (LSTM) deep learning network (Full-BiLSTM) to effectively learn the periodic fMRI brain status changes using both past and future information for each brief time segment. They then fused them to form the final output by taking a dynamic functional connectivity matrix calculated using the sliding window approach as input. [36]. Higher-order functional connectivity was developed by Chen et al., by taking dynamic relations between the brain regions to infer network topology [37]. Yu et al. proposed a novel method using connectivity weighted sparse representation to construct optimal brain functional networks from rs-fMRI data. The method has taken advantage of both Pearson's correlation and sparse representations, which are the two most commonly used brain network modeling approaches. This ensures the construction of more biologically meaningful brain networks by a unified framework that integrates connectivity strength, group structure, and sparsity. Yu et al. used l1-norm regularized linear regression or sparse representation (SR) [38], which learns a linear relationship while BrainNET considers the non-linear relationships. The above mentioned

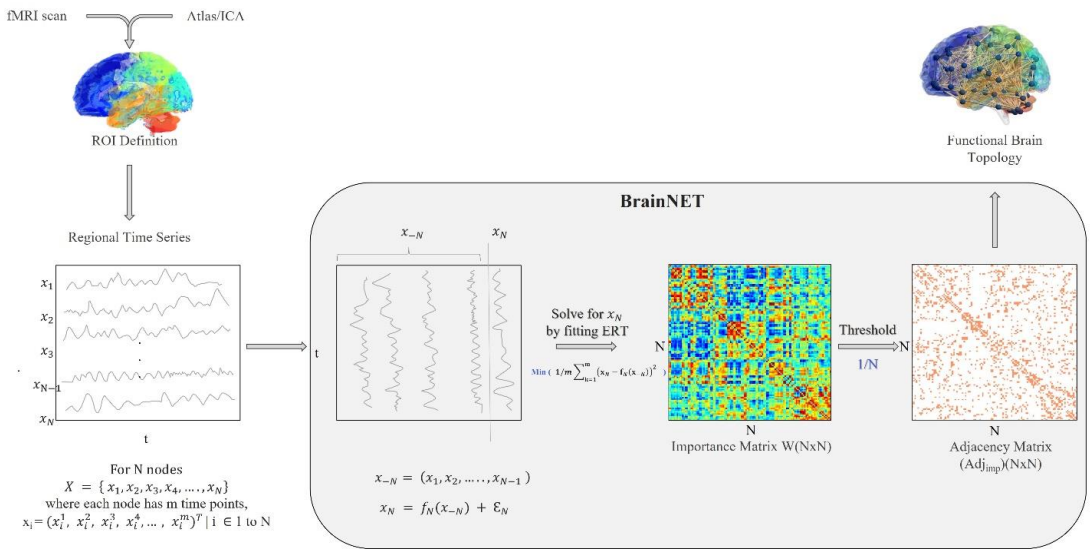


Figure 4: Schematic overview of the BrainNET model. For N nodes in fMRI data (X), each node will have m time points such that $X = \{x_1, x_2, x_3, x_4, \ldots, x_N\}$, *where* x_i *is the vector representation of m time points measured as* $x_i = (x_i^1, x_i^2, x_i^3, x_i^4, \ldots, x_i^m)^T$. *Each node's time series* (x_n) *is predicted from all other nodes time series* (x_{-n}) *using the ERT regressor. Node Importance of each node for predicting the target node are extracted and populated in the importance matrix. The average of the upper and lower triangle of the matrix is thresholded at (1/Num of Nodes) to obtain an adjacency matrix representing the network topology*

methods, including BrainNetClass and GraphVAR focus on using machine learning methods for computer-aided diagnosis by predicting the cognitive metrics or classifying a group of the subjects, whereas BrainNET uses machine learning for inferring the networks directly.

Validation of BrainNET was performed using fMRI simulations with known ground, as well as in real-world ADHD fMRI datasets. In this study, publicly available resting state fMRI simulated data [22] was used to validate BrainNET's ability to infer networks. The real-world performance of BrainNET was then evaluated in a publicly available data set of Attention-deficit/hyperactivity disorder (ADHD). ADHD is one of the most common neurodevelopmental disorders in children with significant socioeconomic and psychological effects [39, 40].It can be difficult to diagnose due to the overlapping nature of symptoms, with resultant diagnostic errors and over-prescribing of medications due to misdiagnosis [41]. ADHD has widespread but often subtle alterations in

multiple brain regions affecting brain function [42, 43]. Neuro Bureau, a collaborative neuroscience forum, has released fully processed open source fMRI data "ADHD-200 preprocessed" from several sites [44, 45] providing an ideal dataset to test the BrainNET model and compare its performance with standard correlation and partial correlation (PC), which is the most widely used methodology to infer brain networks using fMRI data.

2.3 Materials and Methods

2.3.1 Datasets

2.3.1.1 MRI Simulation Data

Open source rs-fMRI simulation data representing brain dynamics was used to validate the BrainNET model [22]. The data were simulated based upon the dynamic causal modeling fMRI forward model, which uses the non-linear balloon model for vascular dynamics, in combination with a neural network model [22]. The open source dataset has 28 simulations; each including simulated data for 50 subjects with a varying number of nodes and several confounders (e.g., shared input between the nodes, varying fMRI session lengths, noise, cyclic connections and hemodynamic lag variability changes). Additional details on the simulations can be found in the original study [22] (Table I).

ADHD data: Preprocessed rs-fMRI data were obtained from the ADHD-200 database (http://fcon 1000.projects.nitrc.org/indi/adhd200/). IRB approval is not required for de-identified data received from an open repository. Seven different sites contributed to the ADHD-200 database for 776 rs-fMRI data acquisitions. The data were preprocessed using the Athena pipeline and was

provided in 3D NifTI format. Additional information on the Athena pipeline and "ADHD 200 preprocessed" data is detailed by Bellec et al [44].

In our study, subjects identified with 'No Naïve medication' status, or questionable quality on rs-fMRI data were excluded. The remaining subjects were age-matched between the groups resulting in 135 Typically Developing Children (TDC) (mean age: 12.00, males: 83), 75 ADHD Inattentive (ADHD-I) (mean age: 11.46, males: 56) and 93 ADHD Combined (ADHD-C) (mean age: 11.86, males: 77) subjects. Mean time series from 116 ROI's in the AAL atlas [46] were extracted using the NILEARN package [47].

2.3.2 BrainNET Model Development

The objective of BrainNET is to infer the connectivity from fMRI data as a network with N different nodes in the brain (i.e., ROI's), where edges between the nodes represent the true functional connectivity between nodes. At each node, there are measurements from m time points $X = \{x_1, x_2, x_3, x_4, \ldots, x_N\}$, where x_i is the vector representation of m time points measured as

$$x_i = (x_i^1, \quad x_i^2, \quad x_i^3, \quad x_i^4, \ldots, \quad x_i^m)^T.$$

Our method assumes that fMRI measurement of BOLD (Blood Oxygen Level Dependent) activation at each node is a function of each of the other nodes' activation with additional random noise.

For the j^{th} node with m time points, a vector can be defined denoting all nodes except the j^{th} node as

$x_{-j} = (x_1, x_2, x_{j-1}, x_{j+1}, \ldots\ldots, x_N)$, then the measurements at the j[th] node can be represented as a function of other nodes as

$$x_j = f_j(x_{-j}) + \varepsilon_j$$

where Ɛj is random noise specific to each nodej. We further assume that function fj () only exploits the data of nodes in x-j that are connected to nodej. The function fj () can be solved in various ways in the context of machine learning. Since the nature of the relationship between different .ROIs in the brain is unknown and expected to be non-linear [48], we choose a tree based ensemble method as it works well with a large number of features with non-linear relationships and is computationally efficient. We utilized Extremely Randomized Trees (ERT), an ensemble algorithm similar to Random Forest, which aggregates several weak learners to form a robust model. ERT uses a random subset of predictors to select divergences in a tree node and then selects the "best split" from this limited number of choices [49]. Finally, outputs from individual trees are averaged to obtain the best overall model [50]. BrainNET infers a network with N different nodes by dividing the problem into N different sub problems, and solving the function fj () for each node independently as illustrated in Figure 4. The steps are listed below:

For j = 1 to N nodes

- Fit the ERT regressor with all the nodes data, except the j[th] node, to find the function f_j that minimizes the following mean squared error:

$$1/m \sum_{k=1}^{m} \left(x_j - f_j(x_{-j})\right)^2$$

- Extract the weight of each node to predict node j,

$$W(j,n) = \begin{cases} w_n & \text{if } n \neq j \\ 0 & \text{if } n = j \end{cases}$$

where w_n is the weight of node to predict node j and n= 1 to N.

- Append the weights values to the Importance matrix

The importance score for each node ($Node_j$) to predict ($Node_i$) is defined as the total decrease in impurity due to splitting the samples based on $Node_j$ [49]. GINI index is used here as the measure of impurity. Let "S" denote a node split in the tree ensemble and let (S_L, S_R) denote it's left and right children nodes. Then, the decrease in impurity ΔImpurity(S) from node split "S" based on $Node_j$ to predict $Node_i$ is defined as

$$\Delta\text{Impurity}(S_{ij}) = \text{Impurity}(S) - (N_L/N_P) * \text{Impurity}(S_L) - (N_R/N_P) * \text{Impurity}(S_R)$$

where, S_L and S_R are left and right splits and N_P, N_L, N_R are number of samples reaching parent, left and right nodes respectively. Let $\mathbb{V}k$ be the number of ensembles, which uses ROI_j for splitting trees. Then, the importance score for $Node_j$

for predicting $Node_i$ is calculated as the average of node impurities across all trees, i.e. Importance of ROI_{ji}

$$I(i,j) = \sum_{G \in \mathbb{V}k} \Delta Impurity\ (S_{ij})/T$$

where T is the number of trees in the ensemble.

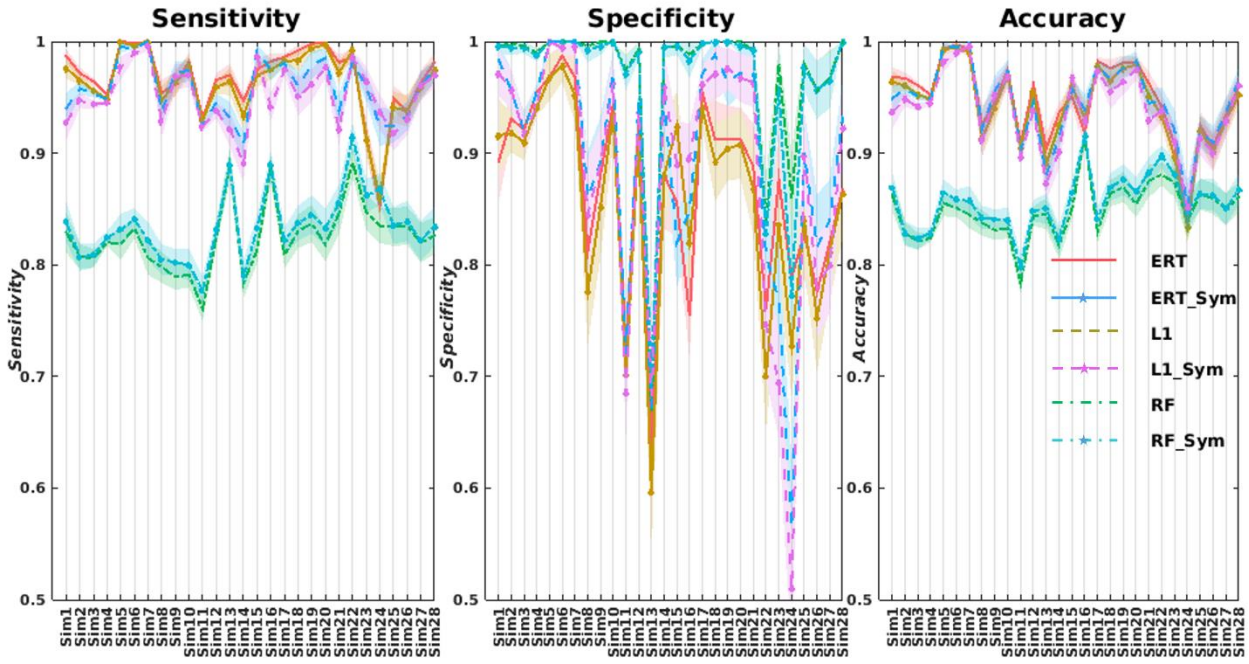


Figure 5: Comparison of BrainNET using ERT, RF and LASSO algorithms to infer brain network topology. Accuracy (Left), Sensitivity (Middle) and Specificity (Right) for BrainNET using Extremely Randomized Trees (ERT), Random Forest (RF), and LASSO (L1) algorithms on symmetrized and non-symmetrized importance matrices. ERT and RF performed similar and better than linear LASSO method. The performance is similar between the symmetrized (ERT_Sym, RF_Sym, RF_Sym) and non-symmetrized matrices for each of the methods.

Table 1: Simulations and corresponding confounders

Simulation	Nodes	fMRI session length (min)	TR	Noise (%)	HRF std dev (s)	Other factors
1	5	10	3	1	0.5	
2	10	10	3	1	0.5	
3	15	10	3	1	0.5	
4	50	10	3	1	0.5	
5	5	60	3	1	0.5	
6	10	60	3	1	0.5	
7	5	250	3	1	0.5	
8	5	10	3	1	0.5	shared inputs
9	5	250	3	1	0.5	shared inputs
10	5	10	3	1	0.5	global mean confound
11	10	10	3	1	0.5	bad ROIs (time series mixed with each other)
12	10	10	3	1	0.5	bad ROIs (new random time series mixed in)
13	5	10	3	1	0.5	backwards connections
14	5	10	3	1	0.5	cyclic connections
15	5	10	3	0.1	0.5	stronger connections
16	5	10	3	1	0.5	more connections
17	10	10	3	0.1	0.5	
18	5	10	3	1	0	
19	5	10	0.25	0.1	0.5	neural lag = 100 ms
20	5	10	0.25	0.1	0	neural lag = 100 ms
21	5	10	3	1	0.5	2-group test
22	5	10	3	0.1	0.5	Non stationary connection strength
23	5	10	3	0.1	0.5	stationary connection strength
24	5	10	3	0.1	0.5	only one strong external input
25	5	5	3	1	0.5	
26	5	2.5	3	1	0.5	
27	5	2.5	3	0.1	0.5	
28	5	5	3	0.1	0.5	

Table 2. Sensitivity, Specificity, Accuracy, and C-sensitivity for each inference method across simulations. Optimum thresholds (with highest performance calculated using simulated data) were used for PC and correlation. The BrainNET threshold was set to [1/number of nodes]. No thresholds were used for C-sensitivity, which is a measure of fraction of true positives that are estimated with a higher connection strength than the 95th percentile of the false positive distribution.

	Sensitivity			**Specificity**			**Accuracy**			**C-Sensitivity**		
Simulation	**Corr**	**PC**	**BN**	**Corr**	**PC**	**BN**	**Corr**	**PC**	**BN**	**Corr**	**PC**	**BN**
1	0.97	1.00	0.99	0.96	0.90	0.89	0.97	0.98	0.97	0.85	0.98	0.90
2	0.94	0.99	0.96	0.99	0.87	0.93	0.94	0.98	0.96	0.48	0.96	0.90
3	0.95	1.00	0.95	0.95	0.80	0.92	0.95	0.98	0.95	0.68	0.92	0.87
4	1.00	1.00	0.92	0.68	0.59	0.95	0.99	0.99	0.92	0.91	0.88	0.90
5	0.96	1.00	1.00	1.00	1.00	0.97	0.97	1.00	0.99	1.00	1.00	1.00
6	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00
7	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	0.99	1.00	1.00	1.00
8	0.75	0.90	0.95	1.00	0.90	0.81	0.80	0.90	0.93	0.06	0.47	0.67
9	0.75	0.90	0.96	1.00	1.00	0.88	0.80	0.92	0.95	0.00	0.20	0.78
10	0.78	0.99	0.98	1.00	1.00	0.94	0.83	0.99	0.97	0.91	1.00	0.94
11	0.80	0.93	0.93	1.00	0.79	0.73	0.82	0.91	0.91	0.18	0.01	0.12
12	0.93	0.99	0.96	0.95	0.77	0.93	0.93	0.96	0.95	0.74	0.84	0.86
13	1.00	1.00	0.95	0.00	0.20	0.86	0.80	0.84	0.93	0.24	0.24	0.72
14	0.79	0.93	0.95	1.00	1.00	0.88	0.82	0.94	0.94	0.10	0.35	0.48
15	0.75	1.00	0.98	1.00	1.00	0.86	0.80	1.00	0.95	0.65	1.00	0.87
16	0.94	1.00	0.98	0.96	0.86	0.75	0.95	0.96	0.92	0.42	0.81	0.81
17	0.92	0.99	0.98	1.00	0.99	0.95	0.93	0.99	0.98	0.69	0.99	0.96
18	0.99	1.00	0.99	1.00	0.94	0.91	0.99	0.99	0.98	1.00	1.00	0.91
19	0.85	1.00	1.00	0.80	0.80	0.91	0.84	0.96	0.98	0.80	0.80	0.96
20	0.85	1.00	1.00	0.80	0.80	0.91	0.84	0.96	0.98	0.80	0.80	0.98
21	0.97	1.00	0.98	0.96	0.90	0.89	0.97	0.98	0.96	0.85	0.98	0.88
22	0.99	0.95	0.99	0.20	0.31	0.84	0.83	0.82	0.96	0.20	0.26	0.81
23	0.75	0.96	0.92	1.00	1.00	0.88	0.80	0.97	0.91	0.60	1.00	0.63
24	0.75	0.90	0.85	1.00	1.00	0.79	0.80	0.92	0.84	0.57	0.82	0.36
25	0.92	0.98	0.94	0.84	0.75	0.83	0.91	0.93	0.92	0.44	0.71	0.72
26	0.98	1.00	0.93	0.35	0.41	0.79	0.86	0.88	0.90	0.55	0.59	0.61
27	0.90	0.96	0.96	0.85	0.80	0.82	0.89	0.93	0.93	0.59	0.76	0.76
28	0.92	0.98	0.98	0.90	0.89	0.87	0.91	0.96	0.95	0.43	0.84	0.88
Average	0.90	0.98	0.96	0.86	0.83	0.88	0.89	0.95	0.95	0.60	0.76	0.80

Table 3. Comparison of Extremely Randomized Trees (ERT), Random forest and LASSO performances in inferring brain network topology using simulation

	Accuracy	Sensitivity	Specificity
Extremely Randomized Trees	94.83	96.71	86.91
Random Forest	94.59	95.88	89.25
LASSO	84.87	82.45	97.00

The importance of each node to predict all other node time series is extracted from the model and an NxN (where N is the number of nodes) importance matrix is generated with the diagonal equal to zero. Each row of the importance matrix represents normalized weights of each node in predicting the target node. The extracted adjacency matrix is affected in two ways. First, due the row-wise normalization, the upper triangular values of the importance matrix are not same as the lower triangle values. We therefore take the average of the upper triangle and the lower triangle of the matrix to make it symmetric to determine the presence of connection between to the nodes.

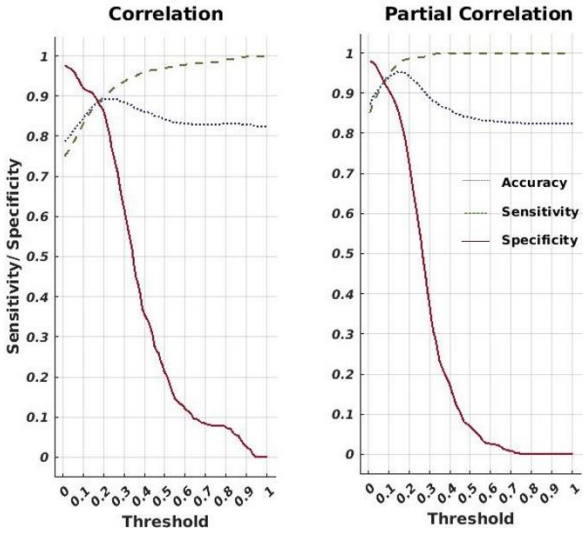


Figure 6: Sensitivity analysis for correlation and partial correlation. Average sensitivity (true positive rate) and specificity across 28 simulations for the correlation and PC method is plotted as a function of threshold ranging between zero to one hundred percent. Optimum threshold is found using simulation ground truth at 20% and 16% for correlation and PC respectively.

This procedure does not allow directionality of the connections to be determined. The comparison between ERT, Random forest and base line LASSO in inferring network topology with the lower and upper triangle, averaged (symmetrized) and non-averaged (non-symmetrized), is provided in Fig.5 (Table.3) Second, again because of the row-wise normalization, the sum of each row in the importance matrix is one. Since the importance values are normalized with respect to number of nodes in the analysis, we used a threshold that is inversely proportional to the number of nodes (i.e., threshold = 1/number of nodes) in the network to produce a final adjacency matrix representing the network topology. The selection of threshold is not based on statistical theory and it is not made to keep the FPR below a nominal level, but it results in a dynamically changing threshold based on the number of nodes in the network.

2.3.3 Analysis

2.3.3.1 Evaluation of inference methods on simulation data

2.3.3.1.1 Evaluation of inference methods on simulation data using C-sensitivity

The network topology was inferred using BrainNET, correlation and PC. The network topology inferred by correlation and PC method may vary drastically based on the values used to threshold connectivity matrix. Hence, we evaluated the ability of the inference methods based on BrainNET, correlation and PC to detect the presence of connection between the nodes in terms of c-sensitivity. C-sensitivity quantifies how well the true positives (TP) are separated from the false positives (FP) by measuring the fraction of TPs that are estimated with a higher connection strength than the 95th percentile of the FP distribution. C-sensitivity is a measure of success in separating true connections from false positive connections and it is calculated by counting

number of true positive above 95th percentile of false positives and then divided by total number of true positives [22].

Effects of simulation parameters such as TR (repetition time), number of nodes, noise, HRF (Hemodynamic Response Function) standard deviation, shared inputs, bad ROI's (Region of Interest), backward, strong and cyclic connections and strong inputs on c-sensitivity of the inference methods were evaluated using mixed-effects model with a random effect for the simulation to control for the effect of the specific generating model. The mixed-effects models were fit across subjects under different simulations to analyze the effects of simulation parameters on C-sensitivity of the inference methods. The parameter estimates from each regression were then summarized across subjects in terms of their effect size [22].

2.3.3.2 Evaluation of inference methods on simulation data using threshold

Thresholding can be applied to suppress spurious connections that may arise from measurement noise and imperfect connectome reconstruction techniques and to potentially improve statistical power and interpretability [21]. However, based on the threshold value, the connection density of each network inferred by correlation and PC may vary from network to network after the threshold has been applied. Using a less stringent lower threshold values results higher false positive values (lower sensitivity) and more stringent threshold results in higher false negatives (lower specificity). This can lead to wide variability in computed graph metrics, as they are typically susceptible to the number of edges in a graph. Identifying an appropriate threshold to infer the underlying brain network topology is critical.

Hence, we evaluated the specificity, sensitivity and accuracy of correlation and PC under varying thresholds. The results show that the network topology inferred using correlation and PC method may vary drastically based on the threshold values (Fig.6). An optimum threshold for correlation methods can

be very difficult to find in real life experimental data. However, given the ground truth for the simulation data we calculated the optimum threshold values for the correlation methods and compared their performance at optimum threshold with BrainNET. The optimal thresholds are defined where the correlation and PC methods performed with the best sensitivity and accuracy. It is important to note that BrainNET is not optimized on this simulation data and the threshold is based on the number of nodes inferred in the network. Specificity, sensitivity and accuracy of correlation and PC at threshold values of 30% ($Corr_{30}$, PC_{30}) and optimum ($Corr_{opt}$, PC_{opt}) values are estimated and compared with BrainNET. We further evaluated the specificity, sensitivity and accuracy of the Corropt and PCopt with BrainNET for each simulation. Similar to the C-sensitivity, we calculated the effect of simulation parameters on sensitivity and specificity of BrainNET, Corropt and PCopt.

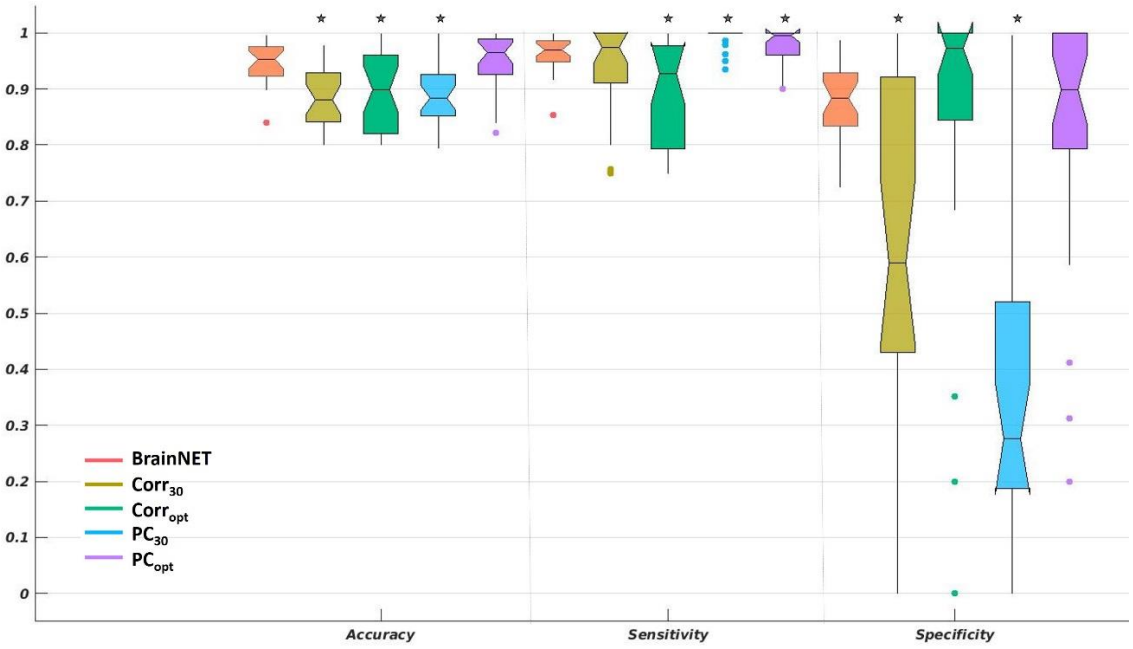


Figure 7: Evaluation of inference methods under varying thresholds. Boxplots of Accuracy (Left), Sensitivity (Middle) and Specificity (Right) across 28 simulations for correlation and PC for optimum and thirty percent threshold ($Corr_{opt}$, Corr30, PC_{opt} and PC_{30}), and BrainNET. '' represents statistically significant differences from BrainNET performance.*

2.3.4 Evaluation of inference methods on ADHD data

BrainNET, correlation and PC was applied to the real-world ADHD data to evaluate whole brain network changes in ADHD subtypes (i.e., ADHD-Combined (ADHD-C), ADHD Inattentive (ADHD-I) compared to Typically Developing Children (TDC). Mean time series from 116 ROI's in the AAL atlas [46] were extracted using the NILEARN package [47]. The BrainNET model was applied to extract an importance matrix for each subject. The importance matrix was then thresholded at 1/number of nodes (e.g., 1/116 for the AAL atlas regions) to obtain an adjacency matrix for each subject (BN). Functional Network connectivity was calculated between the 116 ROIs using correlation and PC. The connectivity matrices are thresholded at a threshold of 20% and 30% ($Corr_{20}$, $Corr_{30}$, PC_{20}, and PC30) (No optimum threshold for real world experimental data). Graph theoretic metrics were extracted using each of these methods for each group. Network differences between the three groups TDC, ADHD-I and ADHD-A were then computed using t-tests on the graph metrics. Site effects and effects of age and handedness were removed using the Combat multi-site harmonization method [51], an effective harmonization technique that removes both unwanted variation associated with the site and preserves biological associations in the data [52].

Graph Metrics: Graph theoretical metrics representing global and local characteristics of network topology were used to compare between the groups in the ADHD dataset. The GRETNA MATLAB toolbox (v2.0,https://www.nitrc.org/projects/gretna/) was used to extract additional graph theoretical metrics including shortest path length, global network efficiency, and betweenness centrality[53]. The Networkx package in python was used to extract the graph theoretical metrics, including Density, Average Clustering Coefficient and Characteristic Path length [54].

Two sample t-tests between groups were performed using the GRETNA toolbox. Bonferroni multiple comparisons correction was applied with statistical significance set at $p < .05$.

Node Metrics: The Nodal Shortest Path Length (NSPL) is defined as the shortest mean distance from a particular node to all other nodes in the graph. Shorter NSPL represents greater integration[15]. The Betweenness Centrality (BC) measures a node's influences in information flow between all other nodes [55]. BC quantifies the influence of a node and is defined as the number of shortest paths passing through it.

Global Metrics: Network Efficiency is a more biologically relevant measure representing the ability of the network to transmit information globally and locally. Networks with high efficiency, both globally and locally, are said to be economic small world networks. [56]. The density of the graph is defined as the ratio of number of connections in the network to the number of possible connections in the network. Average Clustering is the fraction of a node's neighbors that are neighbors of each other. The clustering coefficient of a graph is the average clustering coefficient (ACC) over all nodes in the network. Networks with high clustering coefficient are considered locally efficient networks. Characteristic Path length (CPL) is the average shortest path length between nodes in the graph, with a minimum number of edges that must be traversed to get from one node to another. CPL indicates how easily information can be transferred across the network [15].

2.4 Experimental Results

2.4.1 Simulation Data

2.4.1.1 Evaluation of inference methods using C-sensitivity

BrainNET performed significantly better than correlation ($p<0.001$) and equivalent to partial correlation ($p > 0.05$) methods with c-sensitivity of 79.53%, 59.82% and 75.75% respectively across 28 simulations (Fig.8A) (Table. 2).

The study on effect of simulation parameters on the c-sensitivity of inference methods showed that increasing the number of nodes and session duration doesn't have much effect on any of the inference methods (Fig.8B). Having shared inputs between the nodes affected the correlation and PC method more drastically than BrainNET method. Selection of bad ROIs with mixed time series between them affected all the inference methods negatively, however selection of bad ROIs with randomly random time series mixed between them did not affect the inference methods drastically. Presence of cyclic and backward connection between the nodes affected the correlation and PC methods but not BrainNET. The presence of only one strong input affected the performance of BrainNET method but not the other methods. In summary, BrainNET performance was robust under various confounding factors but prone to selection of inaccurate ROIs with mixed time series between them and networks with only one strong input. Both PC and correlation methods were affected by shared inputs between the nodes, selection of inaccurate ROIs, backward, cyclic and stationary-nonstationary connections between the nodes

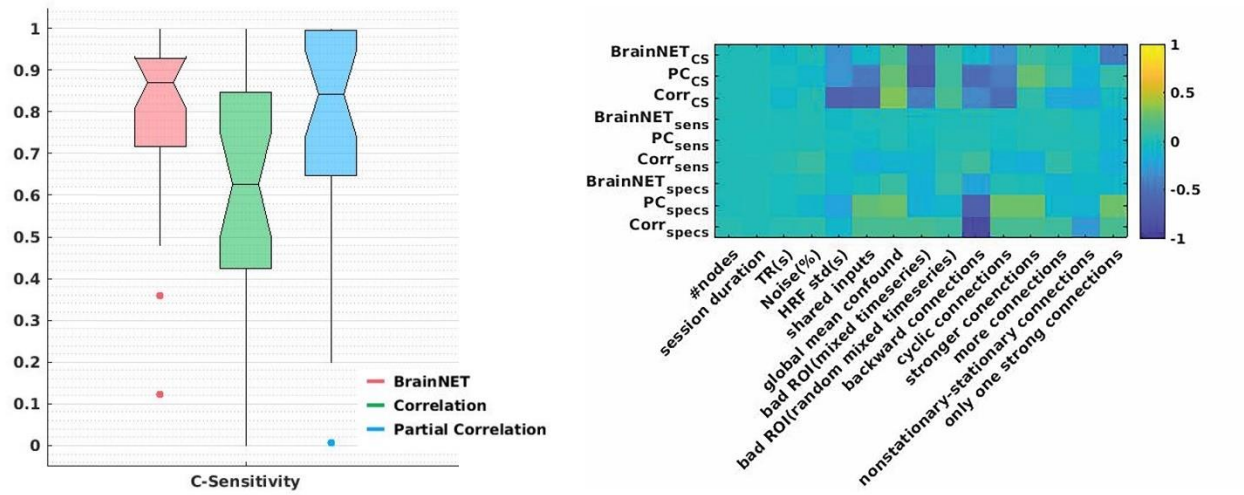


Figure 8: C-sensitivity. Boxplots of c-sensitivity for BrainNET, correlation and partial correlation (PC) (left). The effects of different simulation parameters on the C-sensitivity, sensitivity and specificity of inference methods using mixed effects model. The colorbar represents effect size of each simulation parameter on the c-sensitivity, sensitivity and specificity of inference methods (right).

2.4.1.2 Evaluation of inference methods using threshold

The accuracy, sensitivity, and specificity for each method across all 28 simulated data sets were estimated at thresholds of 30% and at optimum value for correlation and PC, and at threshold of 1/number of nodes for BrainNET (Fig.7). BrainNET achieved higher accuracy and specificity at threshold at 30% compared to the $Corr_{30}$ and PC_{30} method as shown in Fig.7. PC_{opt} achieved slightly higher accuracy than BrainNET across 28 simulations, but no significant difference in terms of specificity and accuracy even at it optimum ($p>0.05$) (Fig.7). As expected the specificity and sensitivity of correlation and PC methods varies with threshold and it will be difficult to find an optimum threshold in a real life dataset. BrainNET showed more robust performance with little variance across the simulation compared to other methods (Fig.9).

The study on effect of simulation parameters on sensitivity and specificity after threshold for BrainNET, PC_{opt} and $Corr_{opt}$, showed that the sensitivity of the inference methods doesn't get affected by the simulation parameters. The specificity of the correlation and PC methods were negatively affected even at their optimum threshold in the presence of backward connections. It is important to note that specificity and sensitivity were calculated after we thresholded the connectivity matrices from each of these methods (1/ number of nodes for BrainNET and optimum threshold for others). After the threshold, BrainNET's performance became robust and consistent across all the simulation parameters.

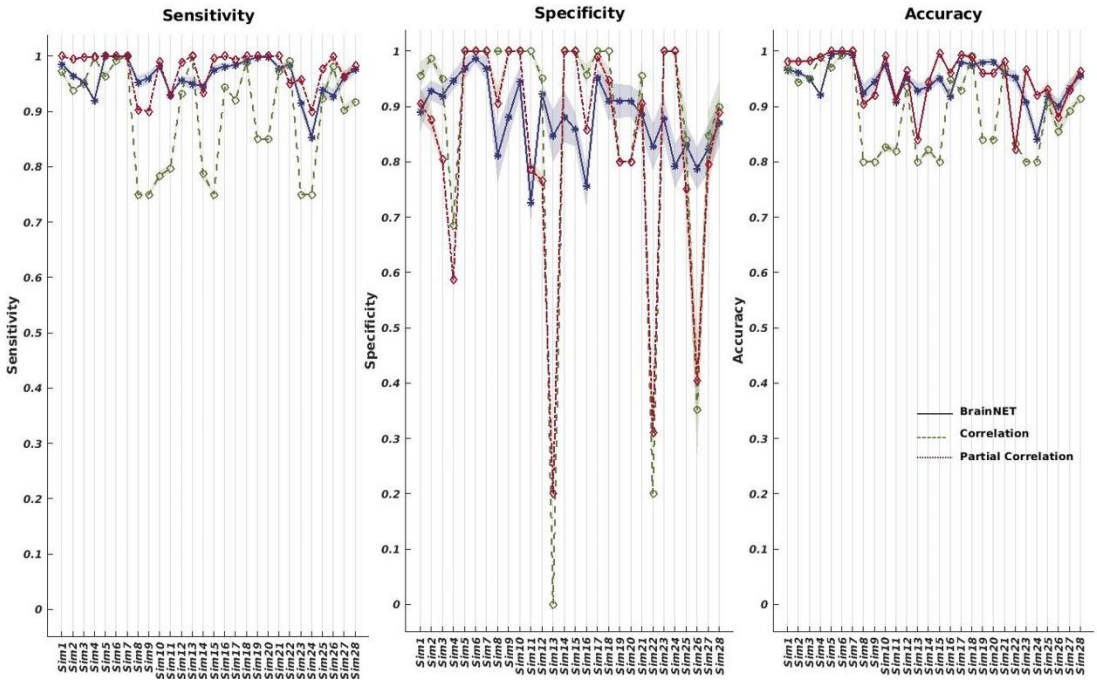


Figure 9: ***Comparison of Correlation ($Corr_{opt}$) and partial correlation (PC_{opt}) at their optimum threshold to BrainNET****. Accuracy (Left), Sensitivity (Middle) and Specificity (Right) for correlation, BrainNET and PC for 28 simulations. Sensitivity, specificity and accuracy are all robust across different simulation cases, while PC and correlation methods show fluctuations even with their optimal threshold for functional connectivity.*

2.4.2 Evaluation of inference methods on ADHD data

BrainNET was able to identify significant changes ($p < 0.05$) in global network efficiency, network density, characteristic path length, betweenness centrality and shortest path in the ADHD data. Correlation and PC was not able to detect significant changes in any of the whole-brain analyses ($Corr_{20}$, $Corr_{30}$, PC_{20}, and PC_{30}).

TDC and ADHD: Statistical analysis of the BrainNET adjacency matrix demonstrated a significant decrease in global Network efficiency, an increase in CPL and an increase in the shortest path length in the right medial temporal gyrus in ADHD compared to TDC (Fig.11A). While the analysis of the correlation adjacency matrix did not show any significant changes, the PC_{30} demonstrated a trending increase in CPL in ADHD compared to TDC ($p=0.07$). Betweenness centrality and Node level local efficiency did not show any changes between the groups in any of the three methods.

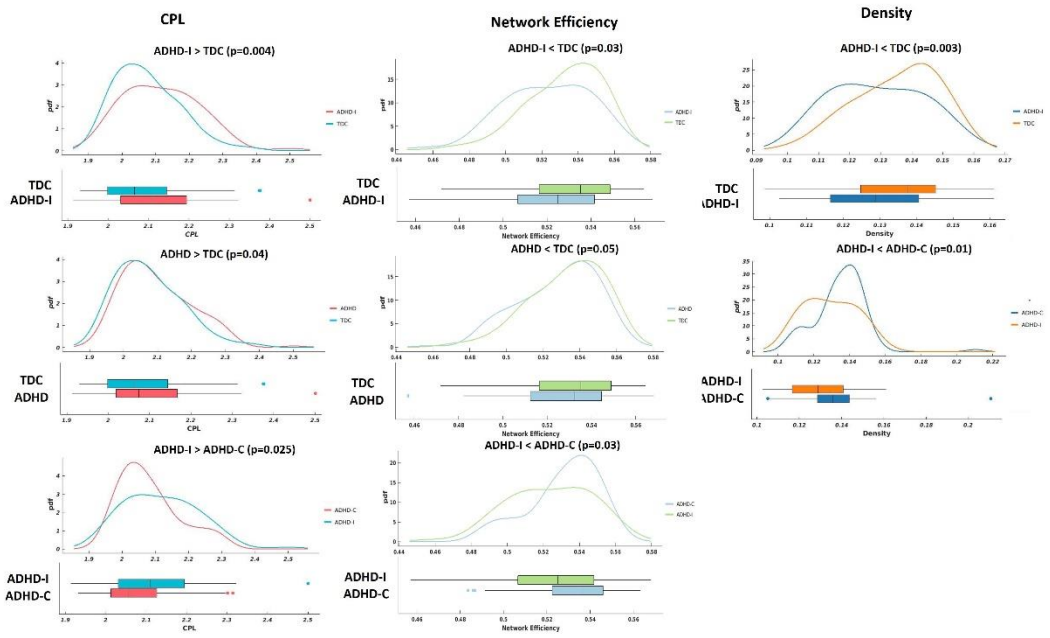


Figure 10: ***Global graph metrics.*** *The probability density functions and boxplots of global graph metrics with significant changes ($p<0.05$) between the groups, ADHD (both ADHD-I and ADHD-C), ADHD-I, ADHD-C and TDC. CPL – Characteristic Path Length.*

TDC and ADHD-I: Statistical analysis of the BrainNET adjacency matrix demonstrated a significant decrease in global network efficiency, a decrease in density, an increase in CPL and an increase in shortest path length in the right superior orbital right, right heschl's gyrus and right medial temporal gyrus nodes in the ADHD-I group compared to TDC (Fig.11A). The correlation method did not show any relation between the groups. No relationship was found in other graph metrics studied in any of the methods.

TDC and ADHD-C: Statistical analysis of the BrainNET adjacency matrix demonstrated a significant decrease in density. No significant relation was found in any other graph metrics for any of the three methods.

ADHD-I and ADHD-C: Statistical analysis of the BrainNET adjacency matrix demonstrated a significant decrease in global network efficiency, a decrease in density, an increase in CPL, an increase in shortest path length of the right olfactory node (Fig.11A). A significant increase in betweenness centrality of the right precuneus node in the ADHD-I group compared to ADHD-C was observed for both BrainNET (Fig.11B). No relationship was found in other graph metrics studied for any of the methods.

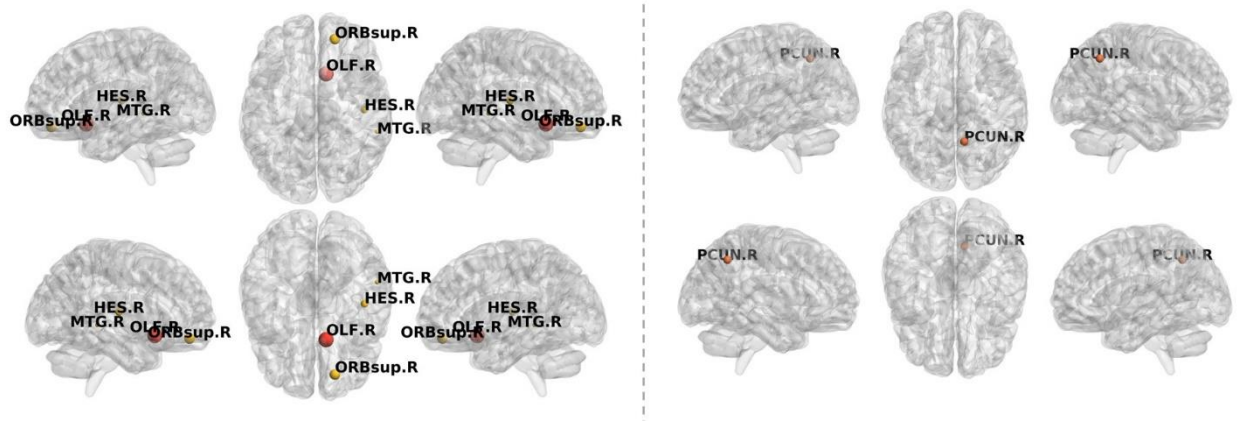


Figure 11: ***Node level graph metrics****. Nodes with significant increases in NSPL in ADHD-I compared to ADHD-C (orange) and in ADHD-I compared to ADHD-C (red) are plotted in the left. Nodes with significant increases in betweenness centrality in ADHD-I compared to ADHD-C are plotted on the right.*

2.5 Discussion

BrainNET was developed to infer brain network topology using ERT [57]. The ERT regressor is used to develop a tree based ensemble model to predict each node's time series from all other node time series. The tree based ensemble methods are ideal for inferring complex functional brain networks as they are efficient in learning non-linear patterns even where there are a large number of features [58]. The importance matrix is then thresholded to generate an adjacency

matrix representing the fMRI topology. The BrainNET model is applicable to both resting-state and task-based fMRI network analysis. It can be easily adapted to datasets with varying session lengths and can be used with different parcellation schemes. A unique feature of the BrainNET approach is that it is implemented at the subject level. It does not need to be trained on big datasets as it infers the network topology based on each individual subject's data.

2.5.1 BrainNET Inference of Network Topology in Simulated fMRI Data

2.5.1.1 Evaluation of inference methods using C-sensitivity

BrainNET demonstrated excellent performance across all the simulations and varying confounders. It achieved significantly higher c-sensitivity than correlation ($p<0.05$) and equivalent to PC ($p=0.38$) (Fig.8A). BrainNET performance remained high in the simulations across varying session lengths, number of nodes, neural lags, cyclic connections, and changing number of connections. BrainNET performed weakest in simulations with one primary strong external source around the network. This causes every node to be highly correlated with other nodes and it becomes very difficult to distinguish direct from indirect connections [22]. It is important to highlight that this kind of one strong external input just for one node is highly unlikely in real life scenarios. BrainNET, similar to PC and correlation methods, was affected by selection of bad ROI's with time series mixed between them. In this simulation, there are 10 nodes, and each node shares a relatively small amount of the other node time series in a proportion of 0.8:0.2. Since the features have shared data between the nodes in this simulation, it limits discrimination of true connectivity between nodes. The leakage of data between nodes can be minimized in fMRI analysis by selecting independent regions using functionally derived parcellation or methods such as ICA.

One concern with this approach is that as the number of nodes increases, the threshold [1/ (number of nodes)] similarly decreases and may result in increased false positives at this low threshold value. The study on effect of number of nodes on the c-sensitivity of BrainNET shows that c-sensitivity of BrainNET does not get affected by the number of nodes (Fig.8B). This can be interpreted that the ability of BrainNET to distinguish between true and false positives increases with the increasing number of nodes and the corresponding lower threshold values doesn't necessarily affect its inference.

Shared inputs between the nodes, can be thought of as distinct sensory inputs that feed into one or more nodes. These shared inputs between the nodes could be deleterious if not modelled [22]. BrainNET is robust to the shared inputs between the nodes, whereas c-sensitivity of PC and correlation are negatively affected (Fig.8B). The performance on varying connection strength over time was tested by simulations of stationary-nonstationary connection strengths between the nodes. BrainNET was least affected by nonstationary connection strengths between the nodes (Fig.8B). The robust performance of BrainNET in simulations with increasing number of nodes, TR, shared inputs, backward, cyclic and non-stationary connections represents a promising aspect of the BrainNET method for inferring brain network topology in real life experimental data (Fig.8B).

2.5.1.2 Evaluation of inference methods using thresholds

In this study we compared the performance of correlation and PC in inferring underlying network topology at optimum threshold values estimated using ground truth ($Corr_{opt}$ and PC_{opt}). We also compared the performance of these methods against BrainNET at 30% threshold ($Corr_{30}$ and

PC_{30}) (Fig.7). BrainNET performed significantly better than PC_{30}, $Corr_{30}$ and $Corr_{opt}$ ($p<0.05$). At its optimum threshold, PC_{opt} performed relatively equivalent to BrainNET in terms of accuracy. However, PC_{opt} showed decreased specificity with increasing number of nodes (sim1-4) and presence of nonstationary, backward connections (sim13 and sim22) (Fig.7). Nonstationary connections represent the varying strengths of connections between nodes which are believed similar to those at the neuronal level and being studied in fMRI. The higher sensitivity and lower specificity of PC_{opt} represents higher numbers of false positive connections which will affect the statistical power of group analysis. The results show that the performance of the PC and correlation method vary under different thresholds and that BrainNET had better performance than these methods even in their optimum (PC_{opt} and $Corr_{opt}$) (Fig.9). The study on effect of number of nodes on the c-sensitivity of BrainNET shows robust performance across all the confounders.

A major strength of the BrainNET approach is that it provides a unique threshold to determine the true network topology. In correlation-based approaches, there is no defined correlation cutoff to determine the true network topology. Instead, multiple approaches are employed, or multiple thresholds applied to generate different networks. Typically, the network cost has been used to define the cutoff value for defining true connections in correlation-based approaches [59]. Multiple costs are then applied to generate multiple instances of the network topology, and analyses are performed to determine the variation in network metrics across these costs, or variation in group differences across thresholds [60]. The BrainNET approach provides a single threshold obviating the need for these imprecise and convoluted thresholding approaches.

2.5.2 Evaluation of inference methods on ADHD Data

Global metrics: Previous studies have shown that ADHD is often associated with changes in functional organization of the brain and lower network efficiencies in ADHD [40, 43]. BrainNET was effective in identifying the subtle changes in the ADHD subjects and supports the notion that the functional organization of brain changes in ADHD (Fig.10) by identifying statistically significant changes in graph metrics between ADHD subjects and typically developing children

Our results demonstrate that there is a decrease in density, network efficiency and an increase in CPL in ADHD compared to TDC. A decrease in density suggests that the number of connections is decreased in ADHD compared to TDC. This can be interpreted as an increase in the cost of wiring in the brain. The increase in CPL and decrease in network efficiency is expected given that there is a decrease in density suggesting that there is increased difficulty in transferring information across the brain in ADHD. The observed abnormalities in global network topology was identified in ADHD-I but not in participants with ADHD-C compared with TDC, however changes between the ADHD-C and ADHD-I were observed. The differential changes observed between the ADHD subtypes may reflect clinical distinctions between the inattentive and combined subtypes of ADHD. Further investigations may shed light on detailed brain-behavior phenotype associations in this neuropsychiatric disorder [61, 62].

Local Metrics:

BrainNET identified increased NSPL in ADHD-I compared to ADHD-C suggesting lesser integration of the prefrontal cortex (PFC) in ADHD-I. The PFC is a part of Default Mode Network (DMN) and plays a crucial role in regulating attention, behavior, and emotion, with the right hemisphere

specialized for behavioral inhibition [63]. The DMN refers to the brain circuitry that includes the medial prefrontal cortex, posterior cingulate, precuneus, and the medial, lateral, and inferior parietal cortices [64]. These results support previous studies demonstrating that ADHD is associated with structural changes and decreased function of the PFC circuits, especially in the right hemisphere [63]. BrainNET also demonstrated that betweenness centrality of the right precuneus, also a part of DMN was increased in ADHD-I compared to ADHD-C group. This suggests increased influence of the precuneus in ADHD-I (Fig.11B). Abnormalities within the DMN have also been found in children in previous studies with ADHD and especially changes in centrality of the right precuneus, which is an important discriminatory feature for classifying ADHD-I and ADHD-C [65] .

Our results also show that the NSPL of the right heschl's gyrus and right medial temporal gyrus is increased in the ADHD-I group compared to TDC. The NSPL of the olfactory cortex was increased in ADHD-I compared to ADHD-C (Fig.11A). Deficits in olfactory function are found in neurodegenerative and neuropsychiatric disorders and represent a topic of interest in ADHD [66]. Increased NSPL was found in the right olfactory region in ADHD-I compared to ADHD-C suggesting lesser integration. Deficits in olfactory ability have been linked to impulsive tendencies within the healthy population and have discriminatory features in identifying people at risk of impulse-control-related problems, supporting the planning of early clinical interventions [67]. Further studies are required to investigate whether functional network topology can be used as a biological marker for early diagnosis, treatment and prognosis of ADHD.

It is important to note that the proposed method measures non-linear relationships while correlation methods measure linear relationships, which may have resulted in the lower performance of correlation in inferring non-linear brain dynamics. Although PC performed relatively similarly to BrainNET in the simulation data, it didn't achieve statistical significance in the ADHD data. This may be due to the false positives identified reducing the statistical power of the analysis. BrainNET can be added to the standard inference methods such as PC and correlation methods, by using a mask derived from BrainNET importance matrix and applying to the correlation matrix. The output from this combined method will have nodes determined by BrainNET, with Pearson correlation values assigned between the connections. This will avoid using arbitrary thresholds, increase the specificity of the standard inference methods by adding non-linearity and allowing analysis of connectivity changes between nodes, which cannot be performed with an adjacency matrix derived only from BrainNET.

Limitations: BrainNET takes relatively longer to infer the adjacency matrix than the correlation method. BrainNET took approximately 3 seconds per subject whereas the correlation and partial correlation method just took 0.001 and 9.3 seconds respectively. Longer running time makes BrainNET challenging to apply for voxel-wise analysis.

2.6 Conclusion

We describe BrainNET, a new network inference method to estimate fMRI connectivity that was adapted from Gene Regulatory methods. We validated the proposed model on ground truth simulation data [22]. BrainNET outperformed Pearson correlation in terms of accuracy and

sensitivity across simulations and various confounders such as the presence of cyclic connections, and even with truncated fMRI sessions of only 2.5 min. We evaluated the performance of BrainNET on the open-source “ADHD 200 preprocessed” data from Neuro Bureau. BrainNET was able to identify significant changes in global graph metrics between ADHD groups and TDC, whereas correlation and PC was unable to find any differences. BrainNET can be used independently or combined with other existing methods as a useful tool to understand network changes and to determine the true network topology of the brain under various conditions and disease states.

2.7 References

1. Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. 2014. 'Machine learning for neuroimaging with scikit-learn', *Frontiers in neuroinformatics*, 8: 14.
2. Achard, Sophie, and Ed Bullmore. 2007a. 'Efficiency and cost of economical brain functional networks', *PLoS computational biology*, 3: e17.
3. Achard, Sophie, and Ed %J PLoS computational biology Bullmore. 2007b. 'Efficiency and cost of economical brain functional networks', 3: e17.
4. Arnsten, Amy FT %J The Journal of pediatrics. 2009. 'The emerging neurobiology of attention deficit hyperactivity disorder: the key role of the prefrontal association cortex', 154: I.
5. Avena-Koenigsberger, Andrea, Bratislav Misic, and Olaf Sporns. 2018. 'Communication dynamics in complex brain networks', *Nature Reviews Neuroscience*, 19: 17.
6. Barber, Anita D, Lisa A Jacobson, Joanna L Wexler, Mary Beth Nebel, Brian S Caffo, James J Pekar, and Stewart H %J Neuroimage: clinical Mostofsky. 2015. 'Connectivity supporting attention in children with attention deficit hyperactivity disorder', 7: 68-81.
7. Bellec, Pierre, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. 2017. 'The neuro bureau ADHD-200 preprocessed repository', *Neuroimage*, 144: 275-86.
8. Breiman, Leo. 2017. *Classification and regression trees* (Routledge).
9. Buckner, Randy L, Jorge Sepulcre, Tanveer Talukdar, Fenna M Krienen, Hesheng Liu, Trey Hedden, Jessica R Andrews-Hanna, Reisa A Sperling, and Keith A Johnson. 2009. 'Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease', *Journal of neuroscience*, 29: 1860-73.
10. Camacho, Diogo M, Katherine M Collins, Rani K Powers, James C Costello, and James J Collins. 2018. 'Next-generation machine learning for biological networks', *Cell*.
11. Chen, Xiaobo, Han Zhang, Yue Gao, Chong-Yaw Wee, Gang Li, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative %J Human brain mapping. 2016. 'High-order resting-state functional connectivity network for MCI classification', 37: 3282-96.
12. Cortese, Samuele, Clare Kelly, Camille Chabernaud, Erika Proal, Adriana Di Martino, Michael P Milham, and F Xavier Castellanos. 2012. 'Toward systems neuroscience of ADHD: a meta-analysis of 55 fMRI studies', *American Journal of Psychiatry*, 169: 1038-55.
13. Crossley, Nicolas A, Andrea Mechelli, Jessica Scott, Francesco Carletti, Peter T Fox, Philip McGuire, and Edward T Bullmore. 2014. 'The hubs of the human connectome are generally implicated in the anatomy of brain disorders', *Brain*, 137: 2382-95.
14. dos Santos Siqueira, Anderson, Biazoli Junior, Claudinei Eduardo, William Edgar Comfort, Luis Augusto Rohde, and João Ricardo Sato. 2014. 'Abnormal functional resting-state networks in ADHD: graph theory and pattern recognition analysis of fMRI data', *BioMed Research International*, 2014.
15. Finkle, Justin D, Jia J Wu, and Neda Bagheri. 2018. 'Windowed Granger causal inference strategy improves discovery of gene regulatory networks', *Proceedings of the National Academy of Sciences*, 115: 2252-57.
16. Fornito, A., A. Zalesky, and M. Breakspear. 2015. 'The connectomics of brain disorders', *Nat Rev Neurosci*, 16: 159-72.
17. Fornito, Alex, Andrew Zalesky, and Edward Bullmore. 2016. *Fundamentals of brain network analysis* (Academic Press).

18. Fortin, Jean-Philippe, Nicholas Cullen, Yvette I Sheline, Warren D Taylor, Irem Aselcioglu, Philip A Cook, Phil Adams, Crystal Cooper, Maurizio Fava, and Patrick J %J Neuroimage McGrath. 2018. 'Harmonization of cortical thickness measurements across scanners and sites', 167: 104-20.
19. Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. 'Extremely randomized trees', *Machine learning*, 63: 3-42.
20. Ghanizadeh, Ahmad, Maryam Bahrani, Ramin Miri, and Ali Sahraian. 2012. 'Smell identification function in children with attention deficit hyperactivity disorder', *Psychiatry investigation*, 9: 150.
21. Hagberg, Aric, Dan Schult, Pieter Swart, D Conway, L Séguin-Charbonneau, C Ellison, B Edwards, and J Torrents. 2013. 'Networkx. High productivity software for complex networks', *Webová strá nka https://networkx. lanl. gov/wiki*.
22. Herman, Aleksandra M, Hugo Critchley, and Theodora %J Scientific reports Duka. 2018. 'Decreased olfactory discrimination is associated with impulsivity in healthy volunteers', 8: 15584.
23. Hilger, Kirsten, and Christian J Fiebach. 2019. 'ADHD symptoms are associated with the modular structure of intrinsic brain networks in a representative sample of healthy adults', *Network Neuroscience*, 3: 567-88.
24. Irrthum, Alexandre, Louis Wehenkel, and Pierre Geurts. 2010. 'Inferring regulatory networks from expression data using tree-based methods', *PloS one*, 5: e12776.
25. Kim, Won Hwa, Nagesh Adluru, Moo K Chung, Ozioma C Okonkwo, Sterling C Johnson, Barbara B Bendlin, and Vikas Singh. 2015. 'Multi-resolution statistical analysis of brain connectivity graphs in preclinical Alzheimer's disease', *Neuroimage*, 118: 103-17.
26. Kim, Won Hwa, Annie M Racine, Nagesh Adluru, Seong Jae Hwang, Kaj Blennow, Henrik Zetterberg, Cynthia M Carlsson, Sanjay Asthana, Rebecca L Koscik, and Sterling C Johnson. 2019. 'Cerebrospinal fluid biomarkers of neurofibrillary tangles and synaptic dysfunction are associated with longitudinal decline in white matter connectivity: A multi-resolution graph analysis', *NeuroImage: Clinical*, 21: 101586.
27. Lin, Pan, Jubao Sun, Gang Yu, Ying Wu, Yong Yang, Meilin Liang, and Xin Liu. 2014. 'Global and local brain network reorganization in attention-deficit/hyperactivity disorder', *Brain imaging and behavior*, 8: 558-69.
28. Milham, Michael P, Damien Fair, Maarten Mennes, and Stewart HMD Mostofsky. 2012. 'The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience', *Frontiers in systems neuroscience*, 6: 62.
29. Murugesan, G., B. Saghafi, E. Davenport, B. Wagner, J. Urban, M. Kelley, D. Jones, A. Powers, C. Whitlow, J. Stitzel, J. Maldjian, and A. Montillo. 2018. 'Single Season Changes in Resting State Network Power and the Connectivity between Regions: Distinguish Head Impact Exposure Level in High School and Youth Football Players', *Proc SPIE Int Soc Opt Eng*, 10575.
30. O'Neill, Thomas J, Elizabeth M Davenport, Gowtham Murugesan, Albert Montillo, and Joseph A Maldjian. 2017. 'Applications of resting state functional mr imaging to traumatic brain injury', *Neuroimaging Clinics*, 27: 685-96.
31. Pellegrini, Enrico, Lucia Ballerini, Maria del C Valdes Hernandez, Francesca M Chappell, Victor González-Castro, Devasuda Anblagan, Samuel Danso, Susana Muñoz Maniega, Dominic Job, and Cyril Pernet. 2018. 'Machine learning of neuroimaging to diagnose cognitive impairment and dementia: a systematic review and comparative analysis', *arXiv preprint arXiv:1804.01961*.
32. Petralia, Francesca, Pei Wang, Jialiang Yang, and Zhidong Tu. 2015. 'Integrative random forest for gene regulatory network inference', *Bioinformatics*, 31: i197-i205.

33. Qian, Xing, Francisco Xavier Castellanos, Lucina Q Uddin, Beatrice Rui Yi Loo, Siwei Liu, Hui Li Koh, Xue Wei Wendy Poh, Daniel Fung, Cuntai Guan, and Tih-Shih %J NeuroImage: Clinical Lee. 2019. 'Large-scale brain functional network topology disruptions underlie symptom heterogeneity in children with attention-deficit/hyperactivity disorder', 21: 101600.
34. Saeed, Fahad. 2018. 'Towards quantifying psychiatric diagnosis using machine learning algorithms and big fMRI data', *Big Data Analytics*, 3: 7.
35. Sidlauskaite, Justina, Karen Caeyenberghs, Edmund Sonuga-Barke, Herbert Roeyers, and Jan R Wiersema. 2015. 'Whole-brain structural topology in adult attention-deficit/hyperactivity disorder: Preserved global–disturbed local network organization', *NeuroImage: Clinical*, 9: 506-12.
36. Smith, S. M., K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. 2011. 'Network modelling methods for FMRI', *Neuroimage*, 54: 875-91.
37. Sporns, Olaf. 2018. 'Graph theory methods: applications in brain networks', *Dialogues in Clinical Neuroscience*, 20: 111.
38. Stam, Cornelis J, and Jaap C Reijneveld. 2007. 'Graph theoretical analysis of complex networks in the brain', *Nonlinear biomedical physics*, 1: 3.
39. Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. 'Bias in random forest variable importance measures: Illustrations, sources and a solution', *BMC bioinformatics*, 8: 25.
40. Supekar, Kaustubh, Mark Musen, and Vinod Menon. 2009. 'Development of large-scale functional brain networks in children', *PLoS biology*, 7: e1000157.
41. Turki, Turki, Jason TL Wang, and Ibrahim Rajikhan. 2016. "Inferring gene regulatory networks by combining supervised and unsupervised methods." In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 140-45. IEEE.
42. Tzourio-Mazoyer, Nathalie, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. 2002. 'Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain', *Neuroimage*, 15: 273-89.
43. Waller, Lea, Anastasia Brovkin, Lena Dorfschmidt, Danilo Bzdok, Henrik Walter, and Johann Daniel %J Journal of neuroscience methods Kruschwitz. 2018. 'GraphVar 2.0: A user-friendly toolbox for machine learning on functional connectivity measures', 308: 21-33.
44. Wang, Jinhui, Xindi Wang, Mingrui Xia, Xuhong Liao, Alan Evans, and Yong %J Frontiers in human neuroscience He. 2015. 'GRETNA: a graph theoretical network analysis toolbox for imaging connectomics', 9: 386.
45. Wang, Jinhui, Xinian Zuo, and Yong %J Frontiers in systems neuroscience He. 2010. 'Graph-based network analysis of resting-state functional MRI', 4: 16.
46. Warren, David E, Jonathan D Power, Joel Bruss, Natalie L Denburg, Eric J Waldron, Haoxin Sun, Steven E Petersen, and Daniel Tranel. 2014. 'Network measures predict neuropsychological outcome after brain injury', *Proceedings of the National Academy of Sciences*, 111: 14247-52.
47. Wehenkel, Marie, Christine Bastin, Christophe Phillips, and Pierre Geurts. 2017. "Tree ensemble methods and parcelling to identify brain areas related to Alzheimer's disease." In *2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 1-4. IEEE.
48. Weyandt, Lisa, Anthony Swentosky, and Bergljot Gyda %J Developmental neuropsychology Gudmundsdottir. 2013. 'Neuroimaging and ADHD: fMRI, PET, DTI findings, and methodological limitations', 38: 211-25.

49. Williams, Nitin, and Richard N Henson. 2018. "Recent advances in functional neuroimaging analysis for cognitive neuroscience." In.: SAGE Publications Sage UK: London, England.
50. Yamashita, Ayumu, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, and Naohiro %J PLoS biology Okada. 2019. 'Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias', 17: e3000042.
51. Yan, Weizheng, Han Zhang, Jing Sui, and Dinggang Shen. 2018. "Deep chronnectome learning via full bidirectional long short-term memory networks for MCI diagnosis." In *International conference on medical image computing and computer-assisted intervention*, 249-57. Springer.
52. Yu, Renping, Han Zhang, Le An, Xiaobo Chen, Zhihui Wei, and Dinggang %J Human brain mapping Shen. 2017. 'Connectivity strength-weighted sparse group representation-based brain network construction for M CI classification', 38: 2370-83.
53. Zaharchuk, G, E Gong, M Wintermark, D Rubin, and CP Langlotz. 2018. 'Deep learning in neuroradiology', *American Journal of Neuroradiology*, 39: 1776-84.
54. Zhou, Zhen, Xiaobo Chen, Yu Zhang, Dan Hu, Lishan Qiao, Renping Yu, Pew-Thian Yap, Gang Pan, Han Zhang, and Dinggang %J Human Brain Mapping Shen. 2020. 'A toolbox for brain network construction and classification (BrainNetClass)'.

3 Resting state fMRI distinguishes subconcussive head impact exposure levels in youth and high school players over a single season of football

Authors: Gowtham Krishnan Murugesan, MS[1,2]., Thomas O Neill, MD[1,2]., Elizabeth M Davenport, PhD[1,2]., Ben Wagner, B.M[1,2]., Jillian E. Urban, PhD[6]., Mireille E. Kelley, MS[6]., Derek A. Jones, MS[6]., Alexander K. Powers, MD[7]., Christopher T. Whitlow, MD., PhD[6,7]., M.H.A, Joel D. Stitzel, PhD[6]., Joseph A. Maldjian[1, 2], M.D and Albert Montillo, PhD[2, 3, 4, 5].

[1,2]Department of Radiology, UT Southwestern, Dallas, TX, USA
[3,4]Department of Bioinformatics, UT Southwestern, Dallas, TX, USA
[4,5]Wake Forest School of Medicine, Winston-Salem, NC, USA

Author contributions: Gowtham Krishnan Murugesan designed the work, analyzed and interpreted the data, and wrote the paper. Dr. Joseph Maldjian, Dr. Albert Montillo, provided expert knowledge and mentorship to develop the method. Ben Wagner contributed in developing MRI analysis. Others, contributed in data collection and to review the paper.

Advances in Knowledge:

- Our study demonstrates the existence of an association between functional changes with subconcussive head impact exposure during a single season of play through a machine learning analysis approach.
- The findings highlight the covert effects of subconcussive head impact exposure and warrant further longitudinal study to understand its physiological and functional consequences over a period of time.

Implication for Patient Care:

- This work demonstrates the ability of machine learning approaches coupled with advanced neuroimaging techniques to identify changes in the functional organization of the brain in a single season of American football.

Summary Statement:

- This study provides additional support to the growing body of evidence that there are detectable changes in the brain from playing a single season of football, even in the absence of clinically diagnosed concussion.

3.1 Abstract

Purpose: To quantify the association of head impact exposure (HIE) with functional changes in the brain measured through resting-state functional MRI (rs-fMRI) using graph theoretical and data-driven methods.

Methods: Anatomical T1 weighted MRI and functional rs-fMRI scans were acquired from 200 athletes immediately before and after a season of American football in an IRB approved study of high school and youth football. During all practices and games, the football players were instrumented with the Head Impact Telemetry (HIT) system, in which an array of accelerometers were mounted inside the football helmet to measure linear acceleration and estimate rotational acceleration of the skull. The combined Risk Weighted cumulative Exposure (RWEcp) for the season was computed for each player using both linear and rotational acceleration. Players with RWEcp higher than two standard deviations from mean exposure are identified as high HIE, and similar data size of age-matched players with lower exposure are identified as low HIE, resulting in 28 high and 34 low HIE players. Neurophysiological relevant networks were extracted using Independent Component Analysis (ICA), and Functional network connectivity (FNC) between the networks is extracted using a Machine learning based network inference method called BrainNET and compared to standard correlation-based connectivity. Percentage changes of global and local graph metrics over a single season of football are extracted from FNC and used as input features to inform four different classifiers to classify HIE of players. The performances of different pipelines are compared, and feature analysis is done on the best performing pipeline. In summary, we developed multiple data-driven and graph theoretical based biomarker extraction pipelines with two different network inference methods to inform four different classifiers to classify HIE of the players.

Results: The pipelines using BrainNET derived connectivity achieved consistently better results than standard correlation methods in all the four classifiers. The combination of BrainNET and the extremely randomized trees classifier achieved top performance with an accuracy of 87.14+/-12.3% (p-value<0.001) in classifying the HIE exposure of the players. Feature analysis using the best performing pipeline suggests that nodal changes in posterior DMN, sensory-motor, and auditory networks provide the highest contribution to the classification power of the models.

Conclusion: Robust performance of multiple pipelines with above chance level predictions demonstrates an association between changes in functional connectivity related to the HIE level in youth and high school football. The current study provides additional support to the growing body of evidence that there are detectable changes in brain health from playing a single season of football, even in the absence of clinically diagnosed concussion.

Keywords: Resting-state, fMRI (rs-fMRI), machine learning, high school, youth, football, head impact exposure, group ICA, subconcussion

3.2 Introduction

In contact sports, such as American football, each player can sustain several hundred subconcussive head impacts in a single season [68]. Understanding the effect of such head impact exposure (HIE) in youth (9-13 years old) and high school (14-19 years old) football players on brain development is of growing concern [69, 70]. Compared to professional and collegiate football players, players at the youth and high school level have received limited attention despite constituting the majority (70%) of all football players [68]. Moreover, these players may be particularly vulnerable to brain injury due to the rapid brain maturation during this age range[70]. Recent resting-state functional MRI (rs-fMRI) studies have provided evidence for changes in the functional connectivity related to subconcussive head impacts in contact sports, including youth populations [71, 72].

Zhu et al. demonstrated the ability of functional connectivity of the default mode network (DMN) to serve as a potential biomarker to monitor dynamic changes in brain function after sports related concussion [73]. Neurophysiological changes in youth football athletes with exposure to subconcussive impacts have also been reported with changes in the resting state DMN [69, 74-77]. Studies have shown that subconcussive impact exposure is associated with more significant disturbances in recognizing emotions and attending to feelings [78]. Previously, high school football athletes with subconcussive head impacts during a season were observed to have decreased functional connectivity in the DMN relative to pre-season measures [75]. Altered DMN functional network connectivity was also reported in 24 active collegiate football players despite receiving only low levels of subconcussive HIE[79, 80] [69]. Functional network changes over a single season were able to distinguish the player's head impact exposure in previous studies, which supports the notion that intrinsic network changes occur as a result of sub-concussive head impacts [26, 79]. Neurocognitive and sensory-motor deficits were reported in military personnel subjected to subconcussive low-level blast events more frequently during training[81]. Recent animal studies have found that low grade cranial impacts in anaesthetized rodent models which were designed to replicate the impact suffered by sports players can result in impaired motor function [82]. Though most of the work focus on studying specific networks such as DMN and sensory-motor, notable studies suggest that subconcussion

can result in spatially heterogeneous changes in local connectivity due to the diffuse nature of head impacts[83].

The purpose of this study is to determine the effect of repetitive subconcussive head impacts on whole-brain network topology using data-driven methods. Our contributions in this work are three-fold: First, we utilized data-driven independent component analysis to identify neurophysiological relevant functional networks. Second, we utilized a machine learning based network topology inference method called BrainNET to infer the functional network connectivity between the extracted functional networks and compared it to the widely used correlation based connectivity. BrainNET is based on Extremely Randomized Trees (ERT) to estimate network connectivity from fMRI data and can generate an adjacency matrix representing brain network topology, without needing arbitrary thresholds as are required in standard approaches. BrainNET is shown to have higher sensitivity than the standard correlation method in identifying true connections between functional nodes and enables to identify subtle functional changes in conditions such as ADHD [84]. Third, we extracted local and global graph metrics from FNC using both BrainNET and the standard correlation method and calculated percentage change of each graph metric over pre and postseason for each player to examine the functional changes between players with low and high HIE over a season of American football using four different machine learning based classification algorithms. Feature analysis of the best performing classifier revealed it had identified regions with functional changes relevant to the HIE level. In summary, we developed multiple data-driven and graph theoretical based biomarker extraction pipelines with two different network inference methods to inform four different classifiers to classify the HIE level of the players.

3.3 Materials and Methods

3.3.1 Study cohort

This study includes 62 football players enrolled in a study of high school and youth football approved by the Wake Forest School of Medicine Institutional Review Board (IRB) [85]. Anatomical T1 weighted MRI and functional rs-fMRI scans were acquired from all subjects before and after the football season. Subjects

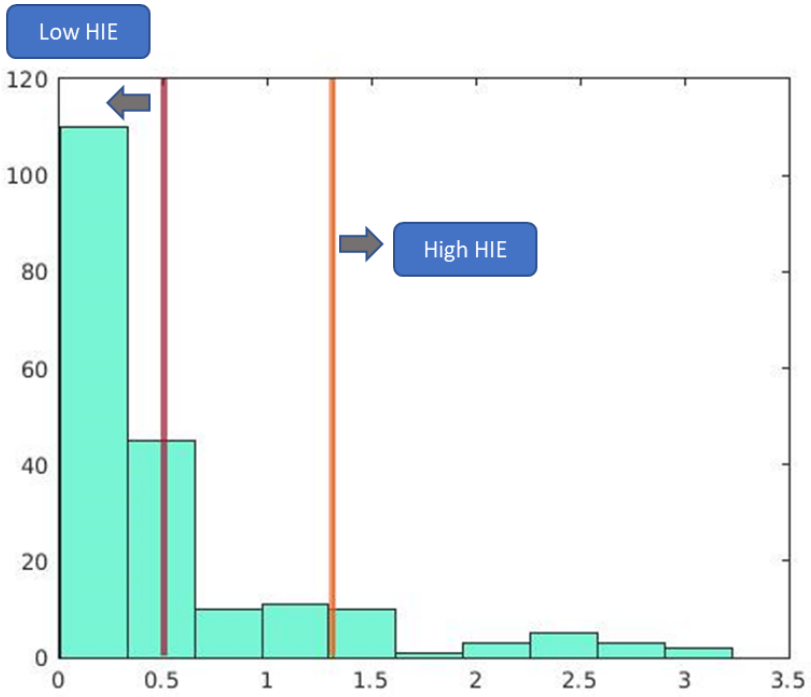


Figure 12: *Distribution of head impact exposure, RWE_{cp}. Subjects with RWEcp above two standard deviation from mean RWEcp are identified as High HIE and age matched players with lower RWEcp of similar data size are identified as Low HIE players. Red and orange line represents mean and two standard deviation from mean*

with a history of concussion, neurological, developmental, or psychiatric disorders were excluded from the study.

3.3.2 Computation of the Head Impact Exposure (HIE) measure

During all practices and games, the football players were instrumented with the Head Impact Telemetry (HIT) system [85, 86], in which an array of accelerometers were mounted inside the football helmet to measure linear acceleration and estimate rotational acceleration of the skull. The Risk Weighted cumulative Exposure (RWE) for the season was computed for each player using linear and rotational acceleration [87]. To compute each RWE metric, the risk of concussion for each impact for each player was calculated using the linear, rotational, and combined probability risk functions previously developed by Rowson et al [88-90]. This biomechanical metric non-linearly weights the severity of each head impact experienced by an athlete over the course of a season using one of three concussion risk functions. The RWE metric evaluated in this study is combined risk-weighted exposure (RWE_{CP}), which accounts for the combined peak resultant linear and rotational accelerations of each impact. Compared to using a discrete measure such as the

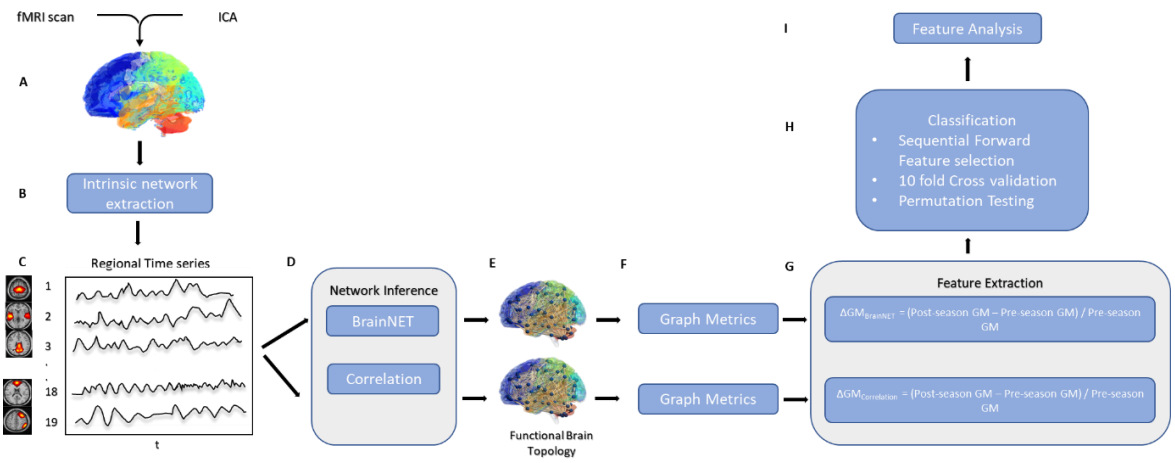


Figure 13: A. Thirty independent components are extracted using independent component analysis B. Nineteen neurophysiological relevant components are selected by expert radiologist (JAM) C. Mean time course for each neurophysiological relevant networks are calculated D. Functional brain network topology is calculated using BrainNET and Correlation methods. E. Five nodal and three global graph metrics are extracted for each subject. F. Percentage change of graph metrics for each player is calculated by considering pre-season metrics as baseline, resulting in total 98 features. G. Four different classifiers are used to classify HIE level with 10 fold cross validation. Permutation testing is done for 1000 times to determine statistical significance of each model. I. Feature analysis is done using top performing model to identify functional changes in brain relevant to HIE level.

number of head impacts a player receives, the RWE metric, a single number representing the cumulative exposure over the season, provides a better quantification of cumulative HIE of the player as it incorporates both the number and magnitude of impacts in a single metric [70].

The players were dichotomized into two groups based on the REWcp. Twenty-eight players with RWEcp two standard deviations above mean RWEcp are considered as High HIE group (RWEcp mean: 1.9055+/-0.6395) and aged matched 34 players with low RWEcp are considered as Low HIE group (RWEcp mean: 0.0472 +/-0.0236). A histogram of the distribution of RWE_{CP} is shown in Fig. 12. In this way, two age-matched groups were formed with a clear separation of HIE levels.

3.3.3 MRI Data Acquisition

MRI data were acquired on a three Tesla Siemens Skyra MRI scanner using a 32-channel human head/neck coil (Siemens Medical, Erlangen, Germany). T1-weighted images were obtained for anatomic correlation using a 3D volumetric Magnetization Prepared Rapid Acquisition Gradient Echo sequence with isotropic resolution of 0.9 mm^3: repetition time (TR) = 1900 msec; echo time (TE) = 2.93 msec; inversion time

(TI) = 900 msec; flip angle = 90 degrees; slices = 176. Six minutes of resting-state fMRI data was obtained using gradient-echo echo planar images (TR = 2, TE = 25 ms, flip angle = 90 degrees, acquisition matrix = 64x64, with a 22 cm FOV, 4 mm slice thickness and 1 mm gap). The participants were instructed to keep their eyes open and fixated on a point. To facilitate group ICA, the fMRI data was preprocessed including motion correction, spatial smoothing and normalization to a common MNI atlas space using SPM8 [91]

3.3.4 Resting State fMRI Analysis and feature extraction

3.3.4.1 Extraction functional connectivity of intrinsic network

After preprocessing the rs-fMRI, ICA was performed to extract thirty (30) independent components using temporal concatenation group (spatial) independent components analysis [92] via the InfoMax ICA algorithm [93]. The number of components is selected such that it allows fine parcellation of intrinsic networks enabling the study of functional connectivity changes between the intrinsic networks such as the DMN, Visual, Language, and Sensory Motor [94]. This approach was applied to the pre- and post-season fMRI. Each extracted component consists of a pair of tensors, including 1) a group spatial map and 2) an activation time course of the spatial map. Back-reconstruction was used to compute subject-specific spatial maps and time courses using the GIFT toolbox [93]. The overall processing pipeline is shown in Fig. 13 Group ICA was applied in order to obtain a consistent set of components for all subjects. ICA outputs components that include both neurophysiological-based components of interest and noise-based, nuisance components. Multiple attributes were employed to separate the noise components from neurophysiological components, including stability across several runs using ICASSO, power ratio, manual inspection as detailed in McKeown et al [95] and a comparison to a functional atlas. In total, 19 components were identified of neurophysiological origin, while 11 components were identified as noise or artifact with the help of an expert neuroradiologist (T.O). The labels and activation maps of each neurophysiological is shown in Figure.16 and 17.

The functional connectivity between networks are calculated using subject specific time course. A mean rs-fMRI time series was extracted from the voxels in each network. Functional network connectivity was

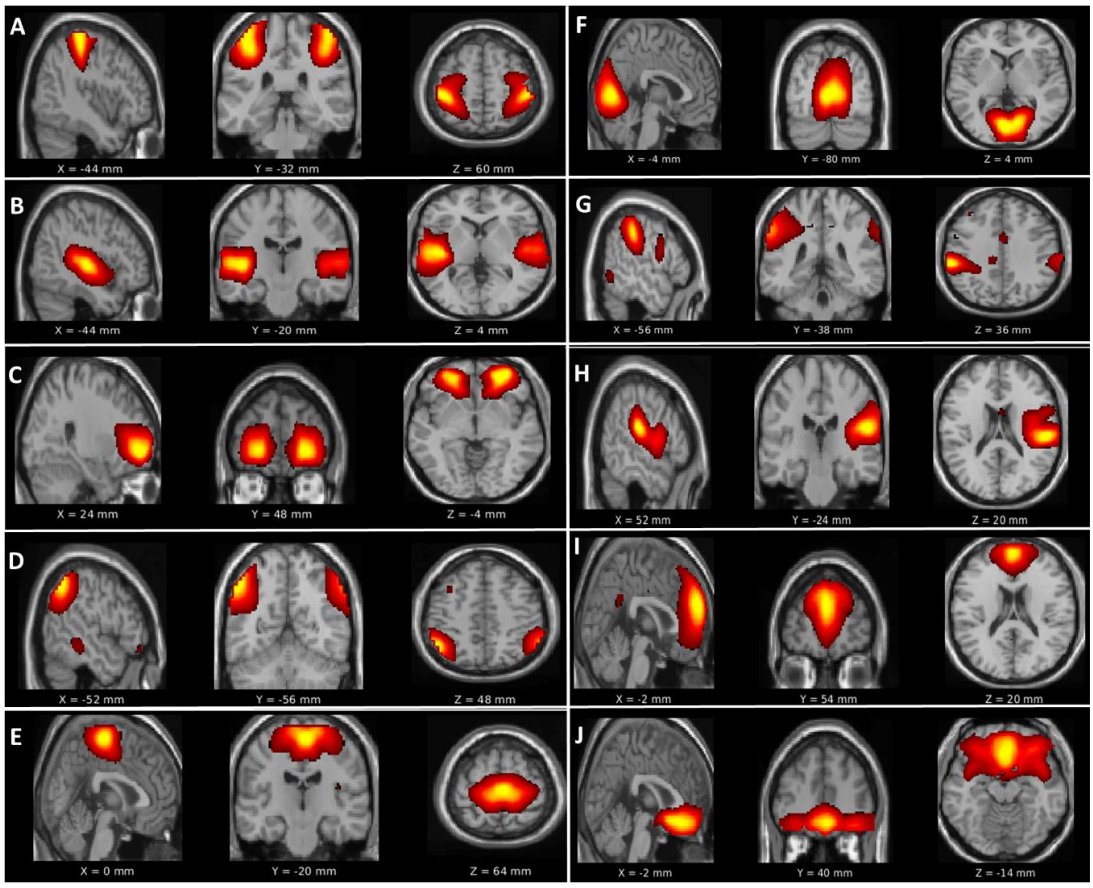


Figure 14:: *Visualization of top 10 feature for HIE level classification. A. Sensory Motor, Auditory, C. Frontal, D. Posterior DMN, E. Sensory Motor, F. Visual Medial, G. Language, H. Auditory, I. Frontal DMN, and J. Frontal*

measured using BrainNET and Pearson's correlation coefficient of the mean time series from each pair of subcomponents. More details on BrainNET can be found in Supplementary Material.

3.3.4.2 Graph Extraction

Graph theoretical metrics representing global and local characteristics of network topology were extracted from both pipelines using BrainNET and correlation methods to compare between the HIE groups. The GRETNA MATLAB toolbox (v2.0,https://www.nitrc.org/projects/gretna/) was used to extract additional graph theoretical metrics including global network efficiency, nodal shortest path length, nodal efficiency, nodal local efficiency, betweenness centrality and degree centrality[53]. The Networkx package in python was used to extract the graph theoretical metrics, including Density and modularity[54]. We utilize

connectivity strength based thresholding for FNC derived using correlation methods. However, such thresholding procedure can lead to different network densities across participants and may confound subsequent between group comparisons. Thus, for correlation based graph metric we corrected for density while calculating graph metrics

In summary, we extracted five nodal metrics as described above for 19 networks (5x19) and three global metrics constituting a total of 98 features per subject. The percentage changes of each graph metrics between post and pre-season (ΔGM) is calculated by subtracting pre-season metrics from post-season metrics and dividing by pre-season metrics. In this way, we performed baseline correction and taking the changes in the metric during the season into account. Two sample t-tests between high and low HIE groups were performed for each graph metrics using the GRETNA toolbox. The nodal comparison is corrected using FDR, and between metrics, comparison is corrected using Bonferroni multiple comparisons correction with statistical significance set at $p < .05$.

Node Metrics: The nodal metrics give complementary information about the organization and the roles of the nodes in the functional connectome. The Nodal Shortest Path Length (NSPL) is defined as the shortest mean distance from a particular node to all other nodes in the graph. Shorter NSPL represents greater integration[15]. The Betweenness Centrality (BC) measures a node's influences in information flow between all other nodes [55]. BC quantifies the influence of a node and is defined as the number of shortest paths passing through it. Degree centrality (DC) calculates the number of direct connections between a given node and the rest of the brain within the entire connectivity matrix of the brain. A node will have a high DC if it has numerous direct connections to other nodes, and thus, DC reflects how much of a node influences the entire brain areas [96]. Nodal clustering coefficient is defined as the number of triangles in the network containing a node divided by the number of connected triples containing the same node [97]. The NCC gives the density of local connections involving a given node and is often used to probe the node's ability to participate in local information integration. Nodal Efficiency assesses the connectedness of the edges among the neighbors of the given node and offers a notion of the network's local robustness to a node's removal [98].

Table 4: Comparison of model performance for pipelines using the BrainNET inference method. Cross-validation mean accuracies (percentages) and corresponding p-values

Classifier	CV accuracy	P-values
Random Forest	75.71+/-11.1	0.0009
ERT	87.14+/-12.3	0.0009
Adaboost	70+/-23.8	0.002
SVC (linear)	65.71+/-19.2	0.03

Global Metrics: Network Efficiency is a more biologically relevant measure representing the ability of the network to transmit information globally and locally. Networks with high efficiency, both globally and locally, are said to be economic small-world networks. [56]. The density of the graph is defined as the ratio of the number of connections in the network to the number of possible connections in the network[15]. Modularity describes the extent to which groups of nodes are connected to the members of their group. Modularity is essentially an index of how cleanly a network can be subdivided with a given partition, with higher values indicating more distinct subnetworks or a greater level of segregation.

3.3.5 Classifier training, evaluation and model selection methodology

Four classifiers were trained using each individual set of features calculated using correlation and BrainNET as inputs (ΔGMBrainNET, ΔGMCorrelation). Historically robust and well performing machine learning classifiers were selected, including Adaboost, linear Support Vector (SV), Random Forest (RF), and Extremely Randomized Trees (ERT) [99]. Features were selected using a sequential forward feature selection algorithm in successive iterations. In the first iteration, all features in the feature space are individually used for classification, and the best performing feature was added to the subset while being removed from the feature space. In each consecutive iteration, individual components of the feature space are added to feature subset, and the best performing feature in combination with previous results is kept for future use. This resulted in 10 features being chosen as the best graph metrics characteristics that distinguish between high HIE and low HIE.

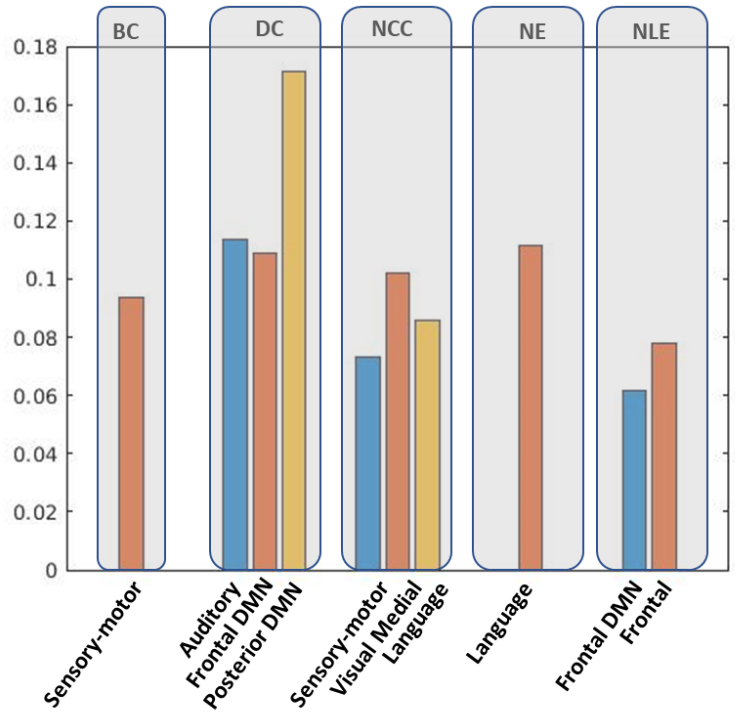


Figure 15:: Visualization of top feature for HIE level classification from top performing pipeline (BrainNET + ERT). Feature analysis shows that changes in nodal degree centrality of posterior DMN being most important feature in identifying HIE level.

The performance of each classifier × feature set combination was systematically evaluated for its ability to distinguish between the high and low impact exposure groups. The subjects were stratified into ten equal parts for both high and low HIE groups, and a 10 fold cross-validation combined with grid search is performed to identify ideal hyperparameters for each classifier (best-classifiers). All the classification metrics were acquired using second 10-fold cross-validation on best classifiers of each algorithm and tested with data split being the same for all algorithms. We minimized the risk of overfitting by using feature selection and using a ten fold cross-validation to ensure that the models are only evaluated on data points that it has not seen before. To assess the reliability of each learning model's accuracy, permutation testing was performed for 1000 times by randomly shuffling labels for each run to calculate p-values. To address the issue of multiple comparisons, we also reported Bonferroni corrected p-values for these classifiers.

3.4 Results

3.4.1 Performance of the classifiers using $\Delta GM_{BrainNET}$ features

Multiple models performed significantly better than chance (chance level accuracy = 45.16%) with p-values<0.006 using graph metrics derived using BrainNET ($\Delta GM_{BrainNET}$). RF and ERT classifiers survived multiple comparisons and performed statistically better than chance level with accuracies 75.714% and 87.14%, respectively, with p-values<0.006 (Table.4). Table 6 also tabulates each graph metrics we studied, and Bonferroni corrected p-value corresponding to the between group difference of that feature. The top 10 most important features activation maps and corresponding feature importances is shown in figure 14 and 15.

3.4.2 Performance of the classifiers using $\Delta GMCorrelation_{features}$

Multiple models performed significantly better than chance (chance level accuracy = 45.16%) with p-values<0.006 using graph metrics derived using correlation ($\Delta GM_{Correlation}$) However, using the standard correlation based connectivity, only ERT classifiers performed significantly better than chance at 78.04% with p<0.006 after multiple comparison correction. (Table. 5) (Figure.18):

Analysis of Selected Features

To further understand the results, we plotted the importance of each selected feature in classifying HIE of players. The results from the top performing pipeline (BrainNET) will be presented in the main body of the article. More details about the performance of the correlation pipeline are given in the supplementary material. The features representing local characteristics of the networks were identified as important features for classification. Measures such as degree centrality of nodal clustering coefficient of frontal DMN, posterior DMN and auditory networks, Betweenness centrality of sensory-motor, the clustering coefficient of visual network and nodal efficiency of language were top five features used in the classification of HIE in players over a single season of football.

3.5 Discussion

In this study, we examined different pipelines, including two different network extraction techniques and using four different classifiers along with sequential forward feature selection. The tested machine learning models identified changes in functional connectivity that discriminate against the HIE level. Multiple models provided robust discrimination between individual players with high and low impact exposure, which underscores the utility of using machine learning to study connectivity changes. This study supports the existence of an association between functional connectivity and HIE during a single season of play. Feature selection helps to avoid overfitting by reducing feature space to be lesser than the number of samples collected. Furthermore, we used 10-fold cross-validation to examine the accuracy of the algorithm for each pipeline, which is shown to be better than the leave-one-out cross-validation used in previous studies [100]. In addition, 10-fold cross-validation may be used as a substitute for having a separate testing set because the model is evaluated on data points it has not seen before [101].

Previous studies use features such as power spectrum density of networks such as DMN and functional connectivity measures extracted using correlation methods between intrinsic networks as features for classifying head impact exposure [26, 79]. In this study, not limited to certain networks, we analyzed whole-brain function changes by utilizing advanced machine learning based BrainNET approach to extract functional network connectivity and extracted global and local graph metrics as features for classification. Graph theoretic measures give the ability to study intermediate and high levels of the organization across the network as a whole, resulting in higher accuracies in classifying HIE of the players than previous studies. The fact that the model identifies nodal metric changes across a season of American football signifies the fact that the diffuse effects of repetitive subconcussion impact on the functional organization of the brain.

3.5.1 Comparison between different pipelines

We extracted graph metrics using two different methods and calculated the percentage changes of each graph metrics across a season of American football ($\Delta GM_{BrainNET}$, $\Delta GM_{Correlation}$). $\Delta GM_{BrainNET}$ features were able to achieve higher classification accuracies with statistical significance. BrainNET measures non-linear relationships and avoids using arbitrary thresholds to obtain functional network topology. As a result, this method is able to correctly identify true connections between the nodes/networks under study with higher sensitivity and hence increases the statistical power of the analysis by reducing false positive connections. Hence, the $\Delta GM_{BrainNET}$ is able to achieve higher accuracies compared to $\Delta GM_{Correlation}$. Among the classifiers tested, ERT is able to achieve better performance than other classifiers RF, Adaboost, and SVC. SVC performed worst may be due to its intrinsic linear nature, making it unable to capture the underlying non-linear dynamics of the brain. The tree-based methods ERT and RF worked better than Adaboost and are based ensemble method as it works well with a large number of features with non-linear relationships and is computationally efficient. Comparing to RF, ERT is robust in the presence of equally important and similar features, such as the graph metrics used in this study. Based on this result, we would suggest BrainNET based network construction pipeline using ERT classifiers for graph theoretical measures analysis of rs-fMRI to identify functional changes in the Brain.

3.5.2 Analysis of the selected features

In this study, we examined changes in graph metrics representing both global and local characteristics of the network topology. Feature analysis identified only changes in graph metrics representing local characteristics of the network topology of different networks to be important in classifying HIE of the players. The nodal clustering coefficient changes of Sensory Motor and Visual Medial network, language, Betweenness Centrality of sensory-motor, Degree centrality of auditory, frontal and posterior DMN, nodal efficiency of the language and nodal local efficiency of frontal DMN and a frontal network were identified as top features important to classify the HIF of the players. Traumatic brain injury and repetitive subconcussion have been shown to affect DMN, and our results support the notion by identifying several parts of the DMN, such as Frontal DMN and Precuneus. The symptoms and signs associated with TBI

include cognitive, motor, mood, and behavioral functions [102]. TBI often affects the visual system; players sustaining a concussion frequently complain of sensitivity to visual stimuli [79]. Our results identified local changes in networks such as visual, sensory-motor, DMN, and language that was known to be affected in TBI; however, our results didn't identify a significant global change [103]. The multitude of networks identified highlights the covert effects of repetitive effects of subconcussive impacts and reflect that continued exposure to repetitive subconcussive impacts may result in substantial neurological and neuropsychological alterations. The identification of local changes in networks that were repeatedly studied in TBI highlights the covert effects of subconcussive HIE and warrants further longitudinal study to understand its physiological and functional consequences over a period of time.

Limitations: There are several limitations to the current study. First, the functional and structural changes we studied are from a single season of American football. Hence the changes may be transient in nature. A long-term longitudinal study with a control group consisting of non-contact sports athletes is needed to understand the evolving functional and structural changes in these young athletes. Second, we have used a smaller dataset for a classification study. In order to identify the difference between high and low impact group, we need to make a clear distinction between the groups using RWEcp, resulting in a smaller subset. Performing further studies with more subjects will certainly add confidence in identifying changes due to repetitive subconcussive impact exposure. Third, although we tested several classifiers and two different methods for FNC extraction, it is entirely possible that several other classifiers and network topology inference methods are better suited for this approach. As an exploratory study to examine functional changes in the brain due to repetitive subconcussive exposure, we tackled the issues by performing feature selection and ten fold cross-validation to achieve as generalized performance as possible. The robust classification accuracies between different pipelines suggest that indeed the proposed pipeline is robust, and there are functional changes in the brain that occurs as a result of subconcussive head impacts.

3.6 Conclusion

In this study, we utilized graph theory and ML and data-driven methods to examine functional changes in the brain over a single season of American football. This study demonstrates an association between changes in functional connectivity related to HIE level in youth and high school football. In particular, this study examined whether a single season of football results in changes that differentiate the HIE level in youth and high school football players in functional connectivity between intrinsic functional networks defined by ICA with rs-fMRI. This result establishes the potential use of these features and machine learning methods to study changes in the intrinsic network connectivity of players with respect to repetitive head impact exposure. The current study provides additional support to the growing body of evidence that there are detectable changes in brain health from playing a single season of football, even in the absence of clinically diagnosed concussion. Long-term risks need to be evaluated in longitudinal studies.

Acknowledgment: Support for this research was provided by NIH R01NS082453 (JAM, JS), and R01NS091602 (JAM, CW, JS). This material is also based upon work supported by the NSF Graduate Research Fellowship under Grant #DGE-0907738. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the NSF.

3.7 References

[1] J. E. Bailes, A. L. Petraglia, B. I. Omalu, E. Nauman, and T. Talavage, "Role of subconcussion in repetitive mild traumatic brain injury," *J Neurosurg,* vol. 119, no. 5, pp. 1235-45, Nov 2013.

[2] B. Johnson, T. Neuberger, M. Gay, M. Hallett, and S. Slobounov, "Effects of subconcussive head trauma on the default mode network of the brain," *J Neurotrauma,* vol. 31, no. 23, pp. 1907-13, Dec 1 2014.

[3] B. R. Cobb *et al.*, "Head impact exposure in youth football: elementary school ages 9-12 years and the effect of practice structure," *Ann Biomed Eng,* vol. 41, no. 12, pp. 2463-73, Dec 2013.

[4] R. Saxena *et al.*, "Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci," *Am J Hum Genet,* vol. 90, no. 3, pp. 410-25, Mar 9 2012.

[5] S. M. Slobounov *et al.*, "The effect of repetitive subconcussive collisions on brain integrity in collegiate football players over a single football season: A multi-modal neuroimaging study," *Neuroimage Clin,* vol. 14, pp. 708-718, 2017.

[6] D. C. Zhu *et al.*, "A potential biomarker in sports-related concussion: brain functional connectivity alteration of the default-mode network measured with longitudinal resting-state fMRI over thirty days," *J Neurotrauma,* vol. 32, no. 5, pp. 327-41, Mar 1 2015.

[7] B. Johnson *et al.*, "Alteration of brain default network in subacute phase of injury in concussed individuals: resting-state fMRI study," *Neuroimage,* vol. 59, no. 1, pp. 511-8, Jan 2 2012.

[8] K. Abbas *et al.*, "Effects of repetitive sub-concussive brain injury on the functional connectivity of Default Mode Network in high school football athletes," *Dev Neuropsychol,* vol. 40, no. 1, pp. 51-6, Jan 2015.

[9] Y. Zhou *et al.*, "Default-mode network disruption in mild traumatic brain injury," *Radiology,* vol. 265, no. 3, pp. 882-92, Dec 2012.

[10] T. J. O'Neill, E. M. Davenport, G. Murugesan, A. Montillo, and J. A. Maldjian, "Applications of Resting State Functional MR Imaging to Traumatic Brain Injury," *Neuroimaging Clin N Am,* vol. 27, no. 4, pp. 685-696, Nov 2017.

[11] C. Gallant and D. Good, "Investigating the relationship between subconcussion and psychiatric symptoms," *Archives of Physical Medicine and Rehabilitation*, vol. 97, no. 10, pp. e65-e66, 2016.

[12] G. Murugesan *et al.*, "Changes in resting state MRI networks from a single season of football distinguishes controls, low, and high head impact exposure," *Proc IEEE Int Symp Biomed Imaging*, vol. 2017, pp. 464-467, Apr 2017.

[13] G. Murugesan *et al.*, "Single Season Changes in Resting State Network Power and the Connectivity between Regions: Distinguish Head Impact Exposure Level in High School and Youth Football Players," *Proc SPIE Int Soc Opt Eng*, vol. 10575, Feb 2018.

[14] F. Haran, J. D. Handy, R. J. Servatius, C. K. Rhea, and J. W. J. A. N. A. Tsao, "Acute neurocognitive deficits in active duty service members following subconcussive blast exposure," pp. 1-13, 2019.

[15] A. P. Lavender *et al.*, "Repeated Long-Term Sub-concussion Impacts Induce Motor Dysfunction in Rats: A Potential Rodent Model," vol. 11, p. 491, 2020.

[16] B. B. Reynolds, A. N. Stanton, S. Soldozy, H. P. Goodkin, M. Wintermark, and T. J. Druzgal, "Investigating the effects of subconcussion on functional connectivity using mass-univariate and multivariate approaches," *Brain imaging and behavior*, vol. 12, no. 5, pp. 1332-1345, 2018.

[17] G. K. Murugesan *et al.*, "BrainNET: Inference of brain network topology using Machine Learning," p. 776641, 2019.

[18] E. M. Davenport *et al.*, "Abnormal white matter integrity related to head impact exposure in a season of high school varsity football," *Journal of neurotrauma*, vol. 31, no. 19, pp. 1617-1624, 2014.

[19] J. J. Crisco *et al.*, "Frequency and location of head impact exposures in individual collegiate football players," *Journal of athletic training*, vol. 45, no. 6, pp. 549-559, 2010.

[20] J. E. Urban *et al.*, "Head impact exposure in youth football: high school ages 14 to 18 years and cumulative impact analysis," *Ann Biomed Eng*, vol. 41, no. 12, pp. 2474-87, Dec 2013.

[21] S. Rowson and S. M. Duma, "Development of the STAR evaluation system for football helmets: integrating player head impact exposure and risk of concussion," *Annals of biomedical engineering*, vol. 39, no. 8, pp. 2130-2140, 2011.

[22] S. Rowson and S. M. Duma, "Brain injury prediction: assessing the combined probability of concussion using linear and rotational head acceleration," *Annals of biomedical engineering*, vol. 41, no. 5, pp. 873-882, 2013.

[23] S. Rowson *et al.*, "Rotational head kinematics in football impacts: an injury risk function for concussion," *Annals of biomedical engineering*, vol. 40, no. 1, pp. 1-13, 2012.

[24] K. J. Friston, "Statistical parametric mapping," 1994.

[25] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. Pekar, "A method for making group inferences from functional MRI data using independent component analysis," *Human brain mapping*, vol. 14, no. 3, pp. 140-151, 2001.

[26] V. Calhoun and T. Adali, "Group ICA of fMRI toolbox (GIFT)," *Online at http://icatb. sourceforge. net*, 2004.

[27] O. Dipasquale, L. Griffanti, M. Clerici, R. Nemni, G. Baselli, and F. Baglio, "High-dimensional ICA analysis detects within-network functional connectivity damage of default-mode and sensory-motor networks in Alzheimer's disease," *Frontiers in human neuroscience*, vol. 9, p. 43, 2015.

[28] M. J. McKeown, L. K. Hansen, and T. J. Sejnowsk, "Independent component analysis of functional MRI: what is signal and what is noise?," *Current opinion in neurobiology*, vol. 13, no. 5, pp. 620-629, 2003.

[29] J. Wang, X. Wang, M. Xia, X. Liao, A. Evans, and Y. J. F. i. h. n. He, "GRETNA: a graph theoretical network analysis toolbox for imaging connectomics," vol. 9, p. 386, 2015.

[30] A. Hagberg *et al.*, "Networkx. High productivity software for complex networks," *Webová strá nka https://networkx. lanl. gov/wiki*, 2013.

[31] O. Sporns, "Graph theory methods: applications in brain networks," *Dialogues in Clinical Neuroscience*, vol. 20, no. 2, p. 111, 2018.

[32] J. Wang, X. Zuo, and Y. J. F. i. s. n. He, "Graph-based network analysis of resting-state functional MRI," vol. 4, p. 16, 2010.

[33] X.-N. Zuo *et al.*, "Network centrality in the human functional connectome," *Cerebral cortex,* vol. 22, no. 8, pp. 1862-1875, 2012.

[34] E. Bullmore and O. Sporns, "The economy of brain network organization," *Nature Reviews Neuroscience,* vol. 13, no. 5, pp. 336-349, 2012.

[35] J. D. Medaglia, "Graph theoretic analysis of resting state functional MR imaging," *Neuroimaging Clinics,* vol. 27, no. 4, pp. 593-607, 2017.

[36] S. Achard and E. J. P. c. b. Bullmore, "Efficiency and cost of economical brain functional networks," vol. 3, no. 2, p. e17, 2007.

[37] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research,* vol. 12, no. Oct, pp. 2825-2830, 2011.

[38] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai,* 1995, vol. 14, no. 2, pp. 1137-1145: Montreal, Canada.

[39] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics,* vol. 7, no. 1, p. 91, 2006.

[40] H. Zetterberg *et al.*, "Head trauma in sports–clinical characteristics, epidemiology and biomarkers," vol. 285, no. 6, pp. 624-634, 2019.

[41] W. H. Thompson, E. P. Thelin, A. Lilja, B.-M. Bellander, and P. Fransson, "Functional resting-state fMRI connectivity correlates with serum levels of the S100B protein in the acute phase of traumatic brain injury," *NeuroImage: Clinical,* vol. 12, pp. 1004-1012, 2016.

[42] C. J. Stam and J. C. Reijneveld, "Graph theoretical analysis of complex networks in the brain," *Nonlinear biomedical physics,* vol. 1, no. 1, p. 3, 2007.

[43] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[44] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics,* vol. 31, no. 12, pp. i197-i205, 2015.

[45] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics,* vol. 8, no. 1, p. 25, 2007.

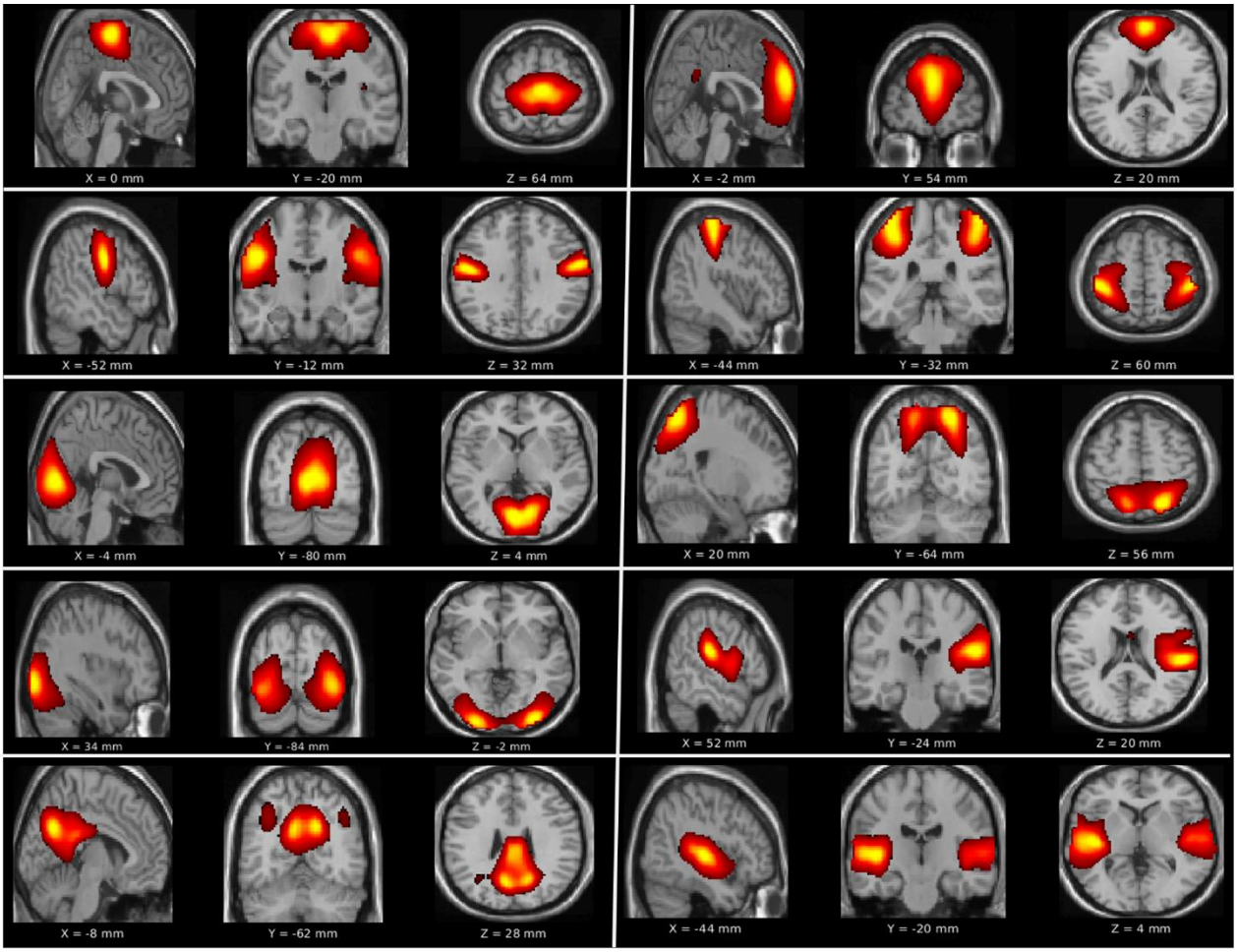


Figure 16: Activation maps of intrinsic components 1-10 extracted using independent component analysis

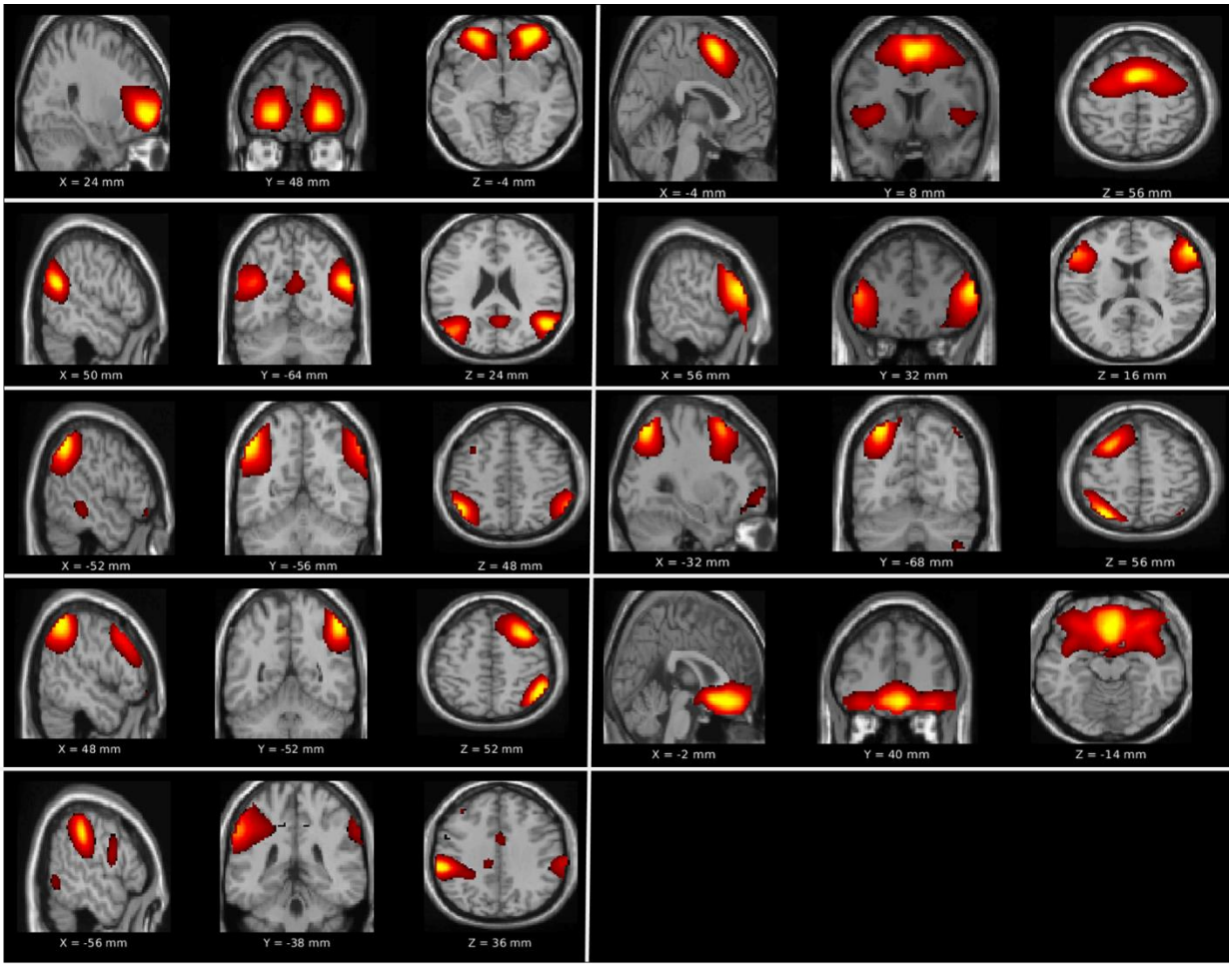


Figure 17: Activation maps of intrinsic components 11-19 extracted using independent component analysis

Table 5: Comparison of model performance for pipelines using Correlation inference method. Cross-validation mean accuracies (percentages) and corresponding p-values

Classifier	CV accuracy	P-values
Random Forest	69.46+/-12.8	0.019
ERT	70.71+/-13.86	0.0009
Adaboost	78.04+/-12.85	0.0009
SVC (linear)	61.96+/-6.08	0.4725

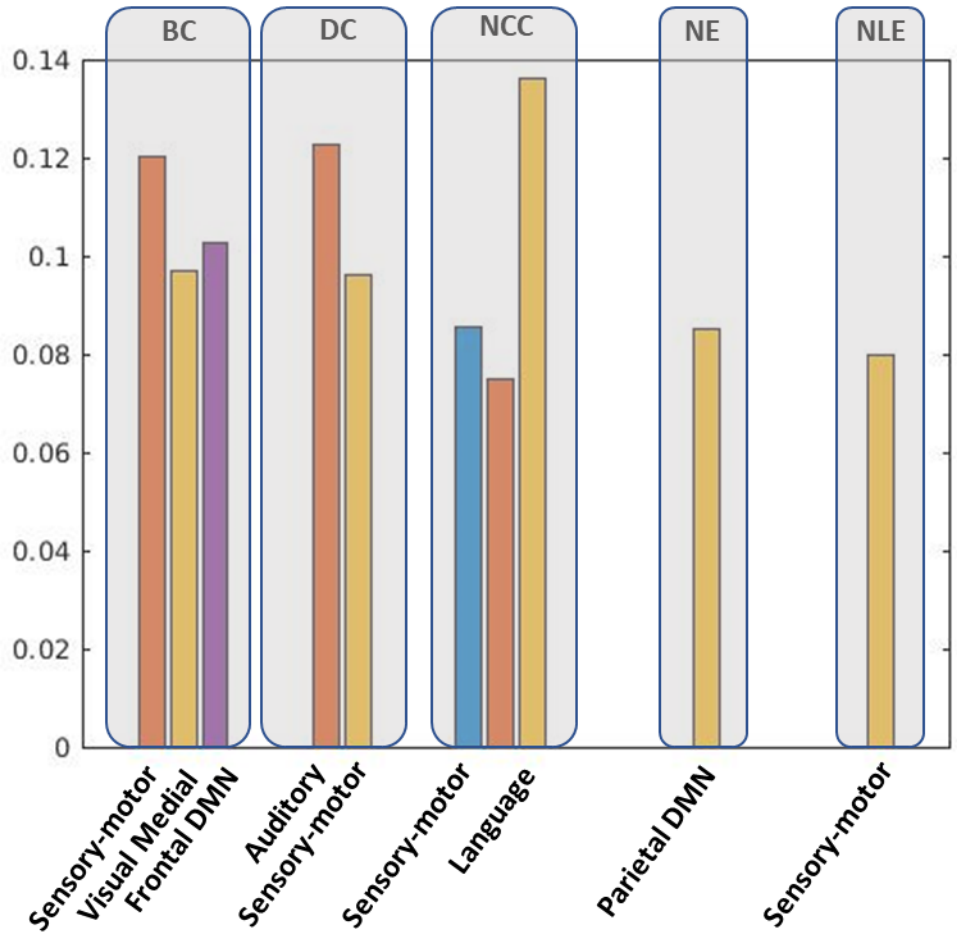


Figure 18: Visualization of top feature for HIE level classification from top performing pipeline using correlation inference method (Correlation + ERT).

*Table **6:** Statistical two sample t-test results between nodal and global graph metrics extracted from nineteen components. Significant p-values after multiple comparisons are highlighted as bold*

Graph Metrics	BrainNET																		
	SM	**SM**	**VM**	**VL**	**Prec.**	**FDMN**	**SM**	**GC**	**Audt**	**Audt**	**Frontal**	**Pa. DMN**	**Pos.DMN**	**FPN**	**GC**	**GC**	**Lang.**	**GC**	**FPN**
BC	0.804	0.086	**0.055**	0.687	0.606	0.121	0.961	0.191	0.168	0.249	0.441	0.267	0.511	0.099	0.071	0.783	0.333	0.676	0.882
DC	0.733	0.557	0.324	0.994	0.765	0.904	0.228	**0.003**	0.160	0.194	0.225	0.722	**0.061**	0.666	0.559	0.234	0.367	0.504	0.782
NCC	0.087	0.086	0.151	0.130	0.882	0.216	0.734	0.832	0.529	0.899	0.932	0.543	0.972	0.373	0.068	0.424	0.955	0.524	**0.062**
NE	0.453	0.543	0.548	0.718	0.916	0.994	0.216	**0.005**	0.147	0.235	0.255	0.916	**0.008**	0.860	0.772	0.276	0.174	0.347	0.793
NLE	0.199	**0.052**	0.270	**0.040**	0.966	0.938	0.794	0.871	0.697	0.548	0.656	0.994	0.567	0.656	0.153	0.404	0.821	0.626	0.095
GE	0.997																		
Mod	0.390																		
Dens.	0.549																		
	Correlation Coefficient																		
BC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	1.000	1.000	1.000	0.983	1.000	1.000	1.000
DC	0.727	0.660	0.403	0.649	0.746	0.612	0.707	0.594	0.673	0.556	0.571	0.703	0.562	0.379	0.738	0.669	0.661	0.487	0.487
NCC	0.912	0.839	0.607	1.000	1.000	0.868	0.908	1.000	0.775	0.922	0.699	0.975	0.924	0.904	0.686	0.894	1.000	1.000	0.682
NE	0.941	0.978	0.660	0.671	0.816	0.837	0.859	0.718	0.953	0.903	1.000	0.924	0.706	0.826	0.942	0.906	0.704	0.764	0.734
NLE	1.000	1.000	0.505	0.826	0.945	0.994	1.000	1.000	0.924	1.000	0.910	1.000	0.975	1.000	0.849	1.000	0.850	0.491	0.538
GE	0.917	0.747	0.824																
Mod																			
Dens.																			

Appendix I.

BrainNET Network Inference Methodology

The objective of BrainNET is to infer the connectivity from fMRI data as a network with N different nodes in the brain (i.e., ROI's), where edges between the nodes represent the true functional connectivity between nodes. At each node, there are measurements from m time points $X = \{x_1, x_2, x_3, x_4, \ldots, x_N\}$, where x_i is the vector representation of m time points measured as

$$x_i = (x_i^1, \quad x_i^2, \quad x_i^3, \quad x_i^4, \ldots, \quad x_i^m)^T.$$

Our method assumes that fMRI measurement of BOLD (Blood Oxygen Level Dependent) activation at each node is a function of each of the other nodes' activation with additional random noise.

For the j^{th} node with m time points, a vector can be defined denoting all nodes except the j^{th} node as

$x_{-j} = (x_1, x_2, x_{j-1}, x_{j+1}, \ldots, x_N)$, then the measurements at the j^{th} node can be represented as a function of other nodes as

$$x_j = f_j(x_{-j}) + \varepsilon_j$$

where ε_j is random noise specific to each $node_j$. We further assume that function f_j () only exploits the data of nodes in x_{-j} that are connected to $node_j$. The function f_j () can be solved in various ways in the context of machine learning. Since the nature of the relationship between different ROIs in the brain is unknown and expected to be non-linear [48], we choose a tree based ensemble method as it works well with a large number of features with non-linear relationships and is computationally efficient. We utilized Extremely Randomized Trees (ERT), an ensemble algorithm similar to Random Forest, which aggregates several weak

learners to form a robust model. ERT uses a random subset of predictors to select divergences in a tree node and then selects the "best split" from this limited number of choices [49]. Finally, outputs from individual trees are averaged to obtain the best overall model [50]. BrainNET infers a network with N different nodes by dividing the problem into N different sub problems, and solving the function f_j () for each node independently. The steps are listed below:

For j = 1 to N nodes

- Fit the ERT regressor with all the nodes data, except the j^{th} node, to find the function f_j that minimizes the following mean squared error:

$$1/m \sum_{k=1}^{m} \left(x_j - f_j(x_{-j})\right)^2$$

- Extract the weight of each node to predict node j,

$$W(j,n) = \begin{cases} w_n & \text{if } n \neq j \\ 0 & \text{if } n = j \end{cases}$$

  where w_n is the weight of node to predict node j and n= 1 to N.

- Append the weights values to the Importance matrix

The importance score for each node ($Node_j$) to predict ($Node_i$) is defined as the total decrease in impurity due to splitting the samples based on $Node_j$ [49]. GINI index is used here as the measure of impurity. Let "S" denote a node split in the tree ensemble and let (S_L, S_R) denote it's left and right children nodes. Then, the decrease in impurity ΔImpurity(S) from node split "S" based on $Node_j$ to predict $Node_i$ is defined as

$$\Delta\text{Impurity}(S_{ij}) = \text{Impurity}(S) - (N_L/N_P) * \text{Impurity}(S_L) - (N_R/N_P) * \text{Impurity}(S_R)$$

where, S_L and S_R are left and right splits and N_P, N_L, N_R are number of samples reaching parent, left and right nodes respectively. Let $\mathbb{V}k$ be the number of ensembles, which uses ROI_j for splitting trees. Then, the importance score for $Node_j$

for predicting $Node_i$ is calculated as the average of node impurities across all trees, i.e. Importance of ROI_{ji}

$$I(i,j) = \sum_{G \in \mathbb{V}k} \Delta\text{Impurity}(S_{ij})/T$$

where T is the number of trees in the ensemble.

Importance values extracted using a typical Random Forest model can be biased in the presence of two or more correlated features since the model will randomly assign importance to any one of the equally important features without any preference [104]. This problem is avoided by using the ERT regressor.

Abbreviations

HIE – Head Impact Exposure

DMN – Default Mode Network

ICA – Independent Component Analysis

FNC – Functional Network Connectivity

RWE – Risk-Weighted Cumulative Exposure

CV – Cross-Validation

FDMN – Frontal DMN

ERT - Extremely Randomized Trees

4 No dose Gadolinium contrast using deep learning

Authors: Gowtham Murugesan, Fang F Yu, Sahil Nalawade, Chandan Ganesh, Ben Wagner, Ananth J Madhuranthakam, Michael Achilleos, Joseph A. Maldjian

Author contributions: Gowtham Krishnan Murugesan conceptualized and designed the work, analyzed and interpreted the data, and wrote the paper. Dr. Joseph Maldjian, Dr. Ananth J Madhuranthakam, Michael Achilleos and Dr. Fang F Yu provided expert knowledge and mentorship to develop the method. Ben Wagner contributed in developing MRI analysis. Chandan Ganesh and Sahil Nalawade contributed to review the paper

4.1 Abstract :

Purpose:

In this study, we used novel deep learning approaches to synthesize T1 post-contrast (T1c) Gadolinium enhancement from non-contrast multi-parametric MR images (T1w, T2w, and FLAIR) in patients with primary brain tumors. We utilized imaging data from 335 subjects in the Brain Tumor Segmentation Challenge (BRATS) 2019 training set for training and validation of the network. A held out a set of 125 subjects from the BRATS 2019 validation dataset was used to test the generalization of the model. A residual inception dense network called T1c-ET was developed and trained to simultaneously predict T1c and segmentation of the enhancing tumor (ET). Two expert neuroradiologists independently scored the synthesized post-contrast images using a 3-point scale, evaluating image quality, motion-artifact suppression, and contrast enhancement against the ground truth T1c images. The 3-point scale was defined as follows: overall image quality and ability to synthesize gadolinium enhancement (1, poor; 3, good; 3, excellent). The predicted T1c images demonstrated structural similarity, PSNR, and NMSE scores of 95.62 37.8357, and 0.0549, respectively. Our model was able to synthesize Gadolinium enhancement in 92.8% of the cases. Inter-rater agreement for predicting contrast enhancement was 87.2% (p-value< 4.0e-05). Rater 1 identified 50, 66, and 9 cases and rater 2 identified 43, 57, and 25 cases (out of the 125 cases) as excellent, good, and poor, respectively. We demonstrate the potential of deep learning methods to synthesize T1c images from non-contrast multi-parametric MRI images. The incorporation of additional training data and MR sequences such as Diffusion-Weighted Imaging may help further generalize the model.

Clinical Relevance statement: This study demonstrates the potential of deep learning to reduce the need for intravenous gadolinium contrast in the evaluation of primary brain tumors.

4.2 Introduction

Structural MRI offers superior soft tissue contrast over other imaging modalities and plays a crucial role in the evaluation of brain tumors by providing information about lesion location, the extent of adjacent tissue involvement, and resultant mass effect upon the surrounding brain parenchyma. The administration of intravenous gadolinium-based contrast agents shorten T1 relaxation times and further increase tissue contrast by accentuating areas where contrast agents have leaked through the blood-brain barrier (BBB) into the interstitial tissues resulting in parenchymal enhancement. This BBB breakdown is a key feature seen in certain tumors such as high-grade gliomas and can serve as a prognostic tool [105].

Gadolinium-based contrast agents (GBCAs) have been used for decades in MR imaging and historically considered safe for patients with normal renal function[105]. It is well-known that there is a risk of nephrogenic systemic fibrosis associated with GBCA administration in patients with renal impairment, particularly at higher doses. However, recent studies have shown gadolinium deposition in tissues throughout the body, including the brain, even in the setting of normal renal function, which raises additional concerns about the long-term safety of these agents [106]. Persistently increased signal intensity on T1-weighted (T1w) MRIs have been reported within the dentate nucleus and globus pallidus following prior injections of both linear and macrocyclic GBCAs.

In lieu of these concerns with Gadolinium toxicity, there has been growing interest in alternative approaches for contrast-enhanced MRI. Examples include manganese-based compounds [107] as well as conceptually different approaches such as chemical exchange saturation transfer (CEST) [108]. Recent developments in deep learning algorithms have shown promise in the field of image synthesis and reconstruction. Gong et al. developed a deep learning method to predict full dose T1w post-contrast images from one-tenth of the GBCA dose [109]. Kleeseik et al. developed a Bayesian deep learning architecture for the prediction of virtual contrast enhancement from non-contrast MR images, including native T1w (nT1w), T2w, FLAIR, DWI, and SWI MR images to predict T1w post-contrast images [110]. Ponnada et al. evaluated whether

deep learning can predict enhancing lesions on MRI scans obtained without the use of contrast material in multiple sclerosis patients.

Our contributions in this work are three-fold. First, we developed a novel deep learning network called Residual-Inception-Dense Network (RIDNet) to demonstrate the feasibility of a deep learning model to synthesize T1 post-contrast images using non-contrast FLAIR, T1w, and T2w images. Second, we utilized a multisite imaging data with different imaging characteristics from different scanners to train the model and evaluated the ability of the model to synthesize Gadolinium enhancement by two experienced raters. Using substantially larger testing data from multiple sites having heterogeneous imaging characteristics helps to determine the generalizability of the model accurately. Third, we analyzed the importance and role of each input MR sequence in predicting the contrast enhancement giving insights on using. The main objective of the study is to investigate the potential of deep learning methods and utility of input MR sequences in predicting enhancing lesions without GBCA administration using an open-source multisite tumor data from BRATS.

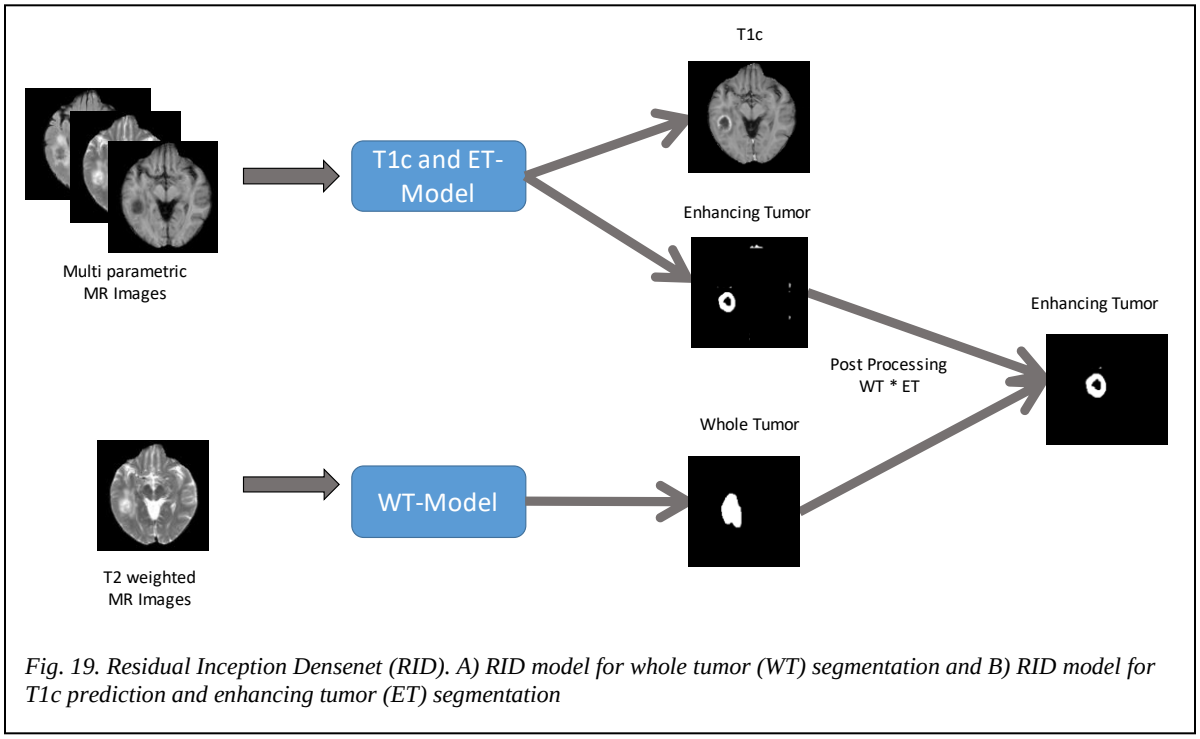


Fig. 19. Residual Inception Densenet (RID). A) RID model for whole tumor (WT) segmentation and B) RID model for T1c prediction and enhancing tumor (ET) segmentation

4.3 Materials and Methods:

4.3.1 Data and Preprocessing

The multimodal Brain Tumor Segmentation Benchmark (BRATS) dataset provides a general platform by outsourcing a unique brain tumor dataset for developing deep learning models [8]. The BRATS 2019 dataset set used in our study was comprised of MRI data of T1, T1c, FLAIR, and T2 from a total of 460 glioma subjects acquired at multiple institutions[111, 112] including contributions from The Cancer Imaging Archive (TCIA), University of Pennsylvania, the University of Alabama in Birmingham, MD Anderson Cancer Center in Texas, the Washington University School of Medicine in St.Louis and the Tata memorial center in India. The imaging characteristics of the BRATS datasets have wide variability in acquisition plane (sagittal/axial/coronal), volume (2D/3D), and variable slice thickness.

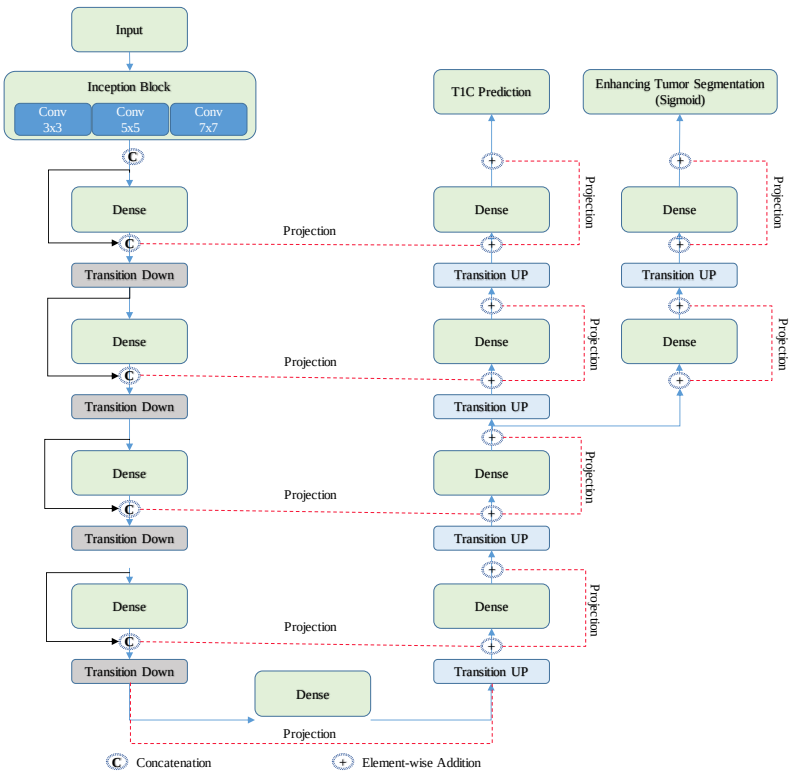


Fig. 20. Residual Inception Densnet (RID) model architecture. RID model for T1c prediction and enhancing tumor (ET) segmentation

All subjects had precontrast T1w, T2w, FLAIR, and post-contrast T1w images. From this set, a single fold

training split of 365 subjects, including 259 HGG subjects and 76 LGG subjects, were used for training, while 125 subjets were held out for testing. The tumor grades for the held out 125 subjects is not made available by BRATS. The training dataset was further randomly split into 300 and 35 for training and validation of the model. Testing on substantially larger testing dataset quantifies the generalizabiltiy of the model accurately.

Data preprocessing

Standard preprocessing steps performed by BRATS include co-registration to an anatomical template[113], resampling to isotropic resolution (1 mm3), and skull-stripping[114]. In addition to that, we performedN4 bias field correction[115] to remove RF inhomogeneity as well as normalizing to zero mean and unit variance.

4.3.2 Network Architecture

4.3.2.1 Model Description

The Residual Inception Dense Network (RID) network was first proposed and developed by Khened et al. for cardiac segmentation. We incorporated our implementation of the RID network with a slight modification in Keras with a Tensorflow backend (Figure.19) (Figure.20). In the DenseNet architecture, the GPU memory footprint increases with the number of feature maps and larger spatial resolution. The skip connections from the down-sampling path to the up-sampling path use element-wise addition in this

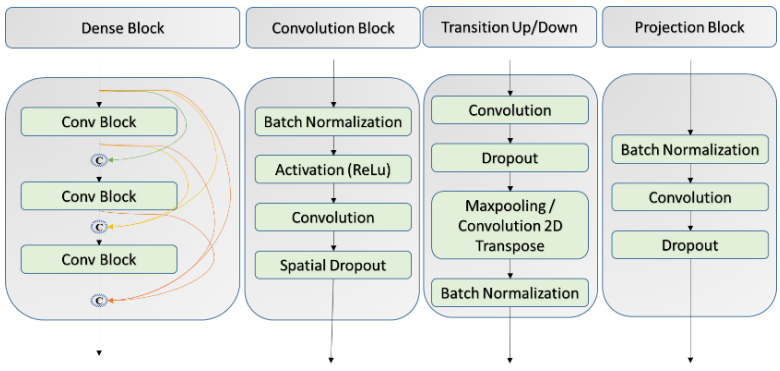


Fig. 21. Building Blocks of Residual Inception Network. From left to right, dense block, convolution block, transition block and projection block

model, instead of the concatenation operation in DenseNet, to mitigate feature map explosion in the up-sampling path. For the skip connections, a projection operation was done using Batch-Norm-1 × 1-convolution-dropout to match the dimensions for element-wise addition (Figure.21). These additions to the DenseNet architecture help in reducing the parameters and the GPU memory footprint without affecting the quality of the segmentation output. In addition to performing dimension reduction, the projection operation helps in learning interactions of cross channel information [116] and faster convergence. Furthermore, the initial layer of the RID networks includes parallel convolutional neural network (CNN) branches similar to the inception module with multiple kernels of varying receptive fields, which help in capturing view-point dependent object variability and learning relations between image structures at multiple scales [117].

4.3.2.2 Model Training

The RID model was trained on 2D input patches of size 64x64x3 extracted from each image slice, with three channels. T1w, T2w, and FLAIR images were concatenated to create three channels of the input. The decoder part of the network was bifurcated to give two outputs: a) T1c prediction and b) Enhancing Tumor (ET) prediction. Linear activation and sigmoid activation was utilized for T1c and ET prediction, respectively. L2 loss assumes the input data set consists of uncorrelated Gaussian signals. This assumption is not always true in real-world data and can result in blurry images. In order to create sharper output images, we propose to optimize the model with Structural Perception Loss for T1c and dice loss for ET segmentation. The Structural Perception loss is a combination of L2, perception, spatial frequency, and structural similarity loss. In each stage, the model is trained until convergence with Adam optimizers with a learning rate of 0.001 using NVIDIA Tesla P40 GPU's.

4.3.3 Structural Perception Loss

The loss function based on the mean squared error between the pixel values of the original and the reconstructed images is the common choice for learning. However, using just MSE (mean squared error; L2 loss) results in blurry image reconstruction [118]. The blurred image reconstruction shows a lack of high

spatial frequency components that represent edges. In addition to L2 loss, to emphasize the high-frequency components, a convolutional layer with a Laplacian filter bank as weights is added to the model and the MSE of features, called the spatial frequency loss (SFL), is computed from the output of each filter. Perceptual and structural similarity (SSIM) based losses were added to improve model performance. We used a pre-trained VGG-16 network to define perceptual loss functions that measure perceptual differences in predicted and ground truth images [119]. The VGG loss network remains fixed during the training process. The model is trained to optimize the combination of all the above losses, which from now we call as structural perception loss. The Structural Perception Loss (SPL) can be represented as follows:

SPL = (1-α). L2+ α. SSIM + α. SFL+ α. Perceptual

4.4 Evaluation and Statistical Analysis

4.4.1 Quantitative Evaluation

Model performance was evaluated by comparing the model prediction to the ground truth. We computed the structural similarity index (SSIM), the peak signal-to-noise ratio (PSNR), normalized mean squared error (NMSE), and the Dice coefficient of the ET mask. The PSNR measures the voxel-wise difference, NMSE captures the L2 loss, Dice evaluates the overlap of predicted ET tumor mask, and SSIM compares nonlocal structural similarity. To evaluate the scores separately for tumor and nontumor regions, we segmented the whole tumor and enhancing tumor regions using the algorithm from Murugesan et al. [120]. Dice score for ET is calculated for whole-brain without any correction, corrected for the whole tumor (after removing predictions outside of whole tumor segmentation), and for ET (after removing predictions outside ET tumor segmentation from Murugesan et al.) to quantify the ability of the model in predicting ET segmentations.

4.4.2 Qualitative Evaluation

To assess the subjective visual quality and the ability of the model to synthesize Gadolinium enhancement, a board-certified neuroradiologist (FY. with eight years of experience) and a radiologist (MA. with six years

of experience) rated the predicted synthesized contrast enhancement maps by comparing them to the ground truth contrast-enhanced T1w scans. For each data set, scores were determined by taking into account general image quality and degree of visual conformity of the tumor region using a 3-point Likert scale comprised of 1 (none), 2 (good), and 3 (excellent) ratings. To determine the inter-rater agreement, the intraclass correlation coefficient was computed using MATLAB. For the differing ratings, the raters agreed on a consensus. This consensus rating was used for correlation with the quantitative scores. The consensus ratings were also dichotomized into low (1) and high (2 – 3) ratings.

Table 7: Quantitative Evaluation

Metrics	Region	T1c-ET
SSIM	Brain	0.913
	Tumor	0.903
	Enhancing Tumor	0.899
NMSE	Brain	0.03
	Tumor	0.01
	Enhancing Tumor	0.009
Dice	Brain	0.32
	Tumor	0.35
	Enhancing Tumor	0.62
PSNR	Brain	64.35
	Tumor	48.99
	Enhancing Tumor	49.93

4.4.3 Importance of the Input MR sequence for prediction

To determine the influence of the individual MRI sequences on the prediction of the T1c image, we tested the trained model by iteratively replacing all voxels within a specific MR sequence with zeros.

4.5 Results

4.5.1 Quantitative Evaluation

The T1c-ET model RID model was tested on 125 held out test subjects. The T1c-ET model synthesized T1c images from multiparametric non-contrast MRI input (FLAIR, T1, and T2). The average PSNR, NMSE, and Structural Similarity index for the whole brain was 64.35, 0.03, and 0.91 (Table.7). The tumor

and ET regions demonstrate lower SSIM and PSNR compared to normal-appearing brain. The Dice coefficient for enhancing tumor on 125 validation subjects was 0.32, 0.35, and 0.62 for uncorrected (whole brain), corrected for the whole tumor, and corrected for ET, respectively. In most cases, the model was able to synthesize T1c images with well-defined enhancing regions as shown in Fig.22 A.

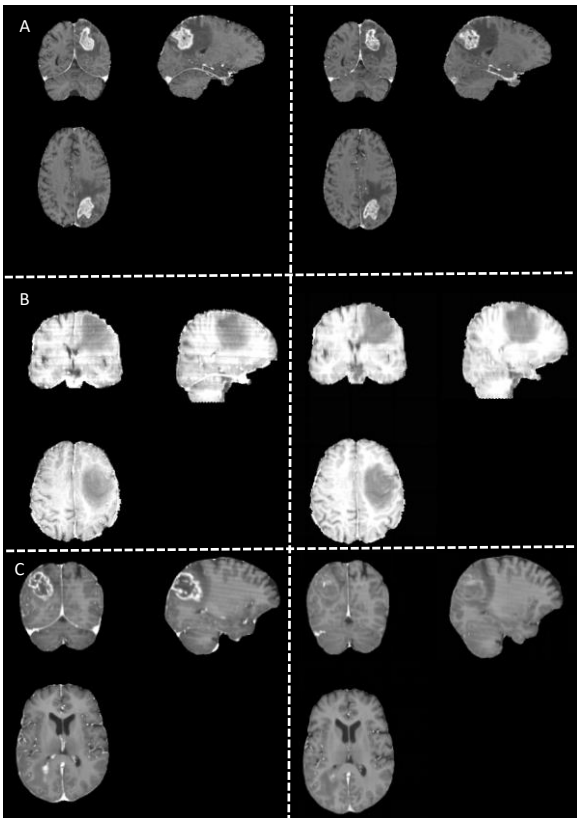


Figure 22: Synthesize T1c images. Ground truth images (left) and Predicted T1c.

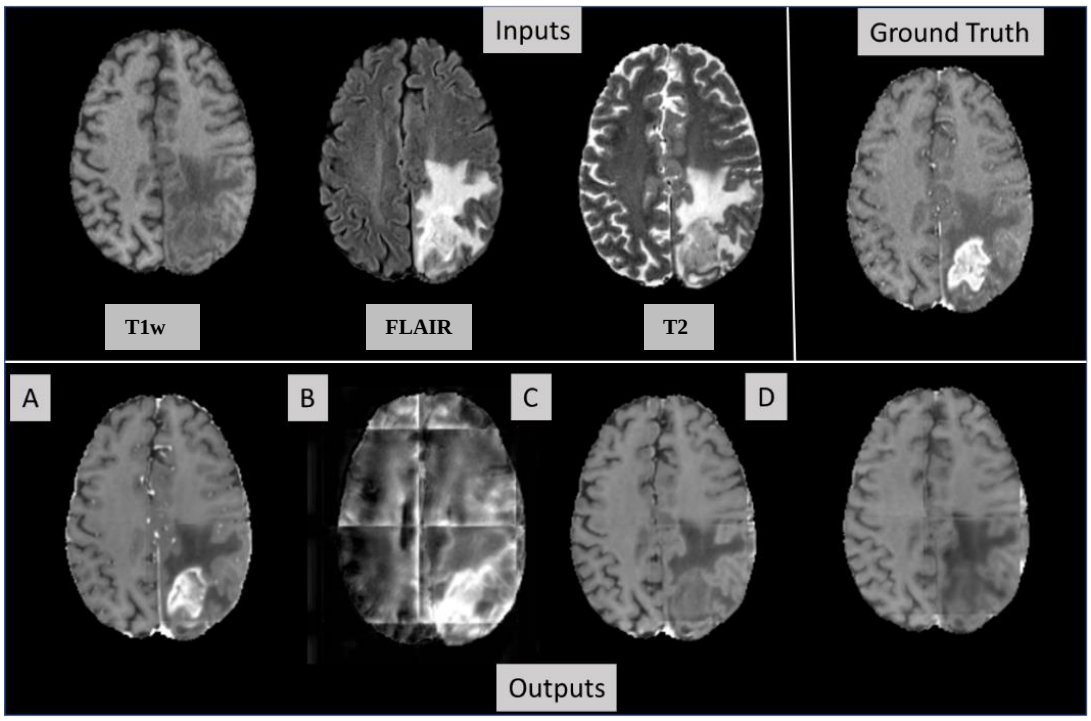


Figure 23: Importance of input sequences. A. Output with all inputs (T1w, FLAIR and T2w) given to the model, B. Output with T1 replaced with zeros in the input, C. Output with FLAIR replaced with zeros in the input, and D. Output with T2 replaced with zeros in the output

4.5.2 Qualitative Evaluation

Examples of representative cases are shown in Figure 22, and animation of real and virtual contrast enhancement next to each other is shown for an entire image volume (Supplemental Digital Content 1) (I will add a video). Comparing the predicted to the ground truth T1w images, 92.8% of the subjective rater scores fell within the good and excellent range. The average rating was 2.18 for enhancing and 2.14 for non-enhancing tumors. The intraclass correlation coefficient of the two radiologists was 0.87, indicating good interrater agreement ($p<1e-4$). In a subset of cases, enhancing regions were not as well captured or missed compared to the ground truth T1c data, an example of which is shown in Figure 22C.

4.5.3 Importance of the input MR sequences for prediction contrast enhancement

By qualitatively examining the outputs of the models by replacing each input sequence with zeros, we were able to determine which sequences are important to predict individual components of the output image. The analysis shows that the T1w image contributes primarily to brain structural information in the predicted

output image. On the other hand, FLAIR and T2w images primarily influence the predicted contrast enhancement (Fig.23).

4.6 Discussion

We demonstrate that the proposed deep learning model was able to synthesize post-contrast T1w images for a majority of primary brain tumor cases using only non-contrast FLAIR, T2w, and T1w images. We tested model performance on a combined dataset of 125 patients scanned using different scanner vendors at multiple institutions. Qualitative and quantitative evaluations of the predicted images show the robust performance of the proposed method for predicting tumor enhancement. In the majority of cases, enhancing and non-enhancing portions of the tumors were correctly predicted. Additionally, the model improved image quality compared to the ground truth contrast-enhanced T1w images.

When compared to earlier work by Gong et al. used low dose Gadolinium-enhanced T1w images to predict full dose T1c images, our results represent an advancement as we were able to synthesize predicted contrast-enhanced images using only non-contrast sequences. Kleesiek et al. developed the Bayesian network to predict post-contrast images using T1, T2, FLAIR, diffusion-weighted imaging (DWI), and susceptibility-weighted imaging (SWI) as a ten channel input and demonstrated the feasibility of synthesizing post-contrast images using non-contrast sequences. Our results further support this approach by demonstrating the successful prediction of enhancement in 91.8% dataset with high ratings from evaluators that we tested. Moreover, we were able to achieve comparable results (and arguably superior in certain quantitative metrics, including PSNR and SSIM) while utilizing fewer sequences (only T1w, T2w, and FLAIR images).

Certain advantages of our strategy included the use of a more diverse dataset through the BRATS dataset. While the two prior studies utilized imaging data from a single institution, BRATS incorporated data from multiple sites and includes imaging of varying acquisition characteristics. While introducing more heterogeneity to the training dataset, this environment also enhances the generalizability of the trained networks. Further, standard preprocessing applied to all the data acquired with different clinical protocols

and various scanners from multiple institutions make this approach useful for generalizing automated approaches across institutions, when differences in hardware and software can significantly alter image representations.

Our qualitative analysis of the importance of input sequences revealed that the FLAIR and T2w images contributed complementary information in predicting ET. This is consistent with the findings from Kleesiek et al., who noted that T2w images were most important to predicting contrast enhancement. FLAIR and T2w images are generally thought of as having greater contrast-to-noise for delineation of pathology compared to T1w images. Changes related to disruption of the blood-brain barrier that leads to contrast enhancement may be better delineated on these sequences, including necrosis and edema. On the other hand, we found that the T1w images contributed information primarily to delineating the overall brain structure. T1w images are often treated as anatomic MR images for their ability to capture anatomic detail, including differentiating between white and grey matter. Of note, although our algorithm utilized only T1w, T2w, and FLAIR images, which could reduce scan time, DWI, and lesser degree SWI (or alternatively, a T2*-weighted gradient echo sequence) are standard sequences for most brain MRI protocols. The incorporation of these additional sequences, particularly DWI, could also further improve the performance of our network.

T1c-ET model failed to predict the Gadolinium enhancement in subjects where one or more of the input sequences has motion corruption and where the tumor has isointensity in both T2 sequences and FLAIR. The failure of the model to predict enhancement in such cases may be due to inadequate representation of such tumors with isoitnensity between FLAIR and T1c in the training set. Another limitation that should be noted is that we used only brain tumor cases for the training of the model. The current method focuses on large tumor lesions and predicting micro brain metastases, and extending it to other body parts needs to be investigated further. Expanding the training dataset to normal subjects with pre and post-contrast imaging could enhance model performance. The current study should be regarded as a clinical feasibility study, and it is not ready for clinical utility. Further studies with larger subjects and different pathologies

should be performed to assess the clinical utility post-contrast synthesize. Perhaps adding low contrast low dose gadolinium contrast along with informative T1w, T2W and FLAIR contrast may increase the efficiency of Gadolinium enhancement model to the clinically applicable methodology.

4.7 Conclusion

We proposed a novel deep learning architecture to synthesize post-contrast enhancement using only non-contrast multiparametric MRI input data. The model demonstrated very good quantitative and qualitative performance in a substantially larger and heterogeneous testing data and showed that the prediction of gadolinium enhancement might be feasible in the near future. FLAIR and T2w images are found to have complementary information in predicting Gadolinium enhancement by performing feature analysis. Further studies in larger patient collectives with varying neurological diseases are needed to assess the clinical practicability of this novel approach.

4.8 References

[1] J. E. Villanueva-Meyer, M. C. Mabray, and S. J. N. Cha, "Current clinical brain tumor imaging," vol. 81, no. 3, pp. 397-415, 2017.

[2] N. Hoggard and G. H. J. T. B. j. o. r. Roditi, "T 1 hyperintensity on brain imaging subsequent to gadolinium-based contrast agent administration: what do we know about intracranial gadolinium deposition?," vol. 90, no. 1069, p. 20160590, 2017.

[3] J. Wang *et al.*, "Manganese-Based Contrast Agents for Magnetic Resonance Imaging of Liver Tumors: Structure–Activity Relationships and Lead Candidate Evaluation," vol. 61, no. 19, pp. 8811-8824, 2018.

[4] S. Viswanathan, Z. Kovacs, K. N. Green, S. J. Ratnakar, and A. D. J. C. r. Sherry, "Alternatives to gadolinium-based metal chelates for magnetic resonance imaging," vol. 110, no. 5, pp. 2960-3018, 2010.

[5] E. Gong, J. M. Pauly, M. Wintermark, and G. J. J. o. M. R. I. Zaharchuk, "Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI," vol. 48, no. 2, pp. 330-340, 2018.

[6] J. Kleesiek *et al.*, "Can virtual contrast enhancement in brain MRI replace gadolinium?: a feasibility study," vol. 54, no. 10, pp. 653-660, 2019.

[7] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993-2024, 2014.

[8] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," vol. 34, no. 10, pp. 1993-2024, 2014.

[9] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," vol. 4, p. 170117, 2017.

[10] T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. J. H. b. m. Pfefferbaum, "The SRI24 multichannel atlas of normal adult human brain structure," vol. 31, no. 5, pp. 798-819, 2010.

[11] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018.

[12] N. J. Tustison *et al.*, "Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements," vol. 99, pp. 166-179, 2014.

[13] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[14] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[15] N. Ichimura, "Spatial Frequency Loss for Learning Convolutional Autoencoders," *arXiv preprint arXiv:1806.02336*, 2018.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, 2016: Springer, pp. 694-711.

[17] G. K. Murugesan *et al.*, "Multidimensional and Multiresolution Ensemble Networks for Brain Tumor Segmentation," p. 760124, 2019.

5 Multidimensional and Multiresolution Ensemble Networks for Brain Tumor Segmentation

Published: Murugesan, Gowtham Krishnan, et al. "Multidimensional and Multiresolution Ensemble Networks for Brain Tumor Segmentation." International MICCAI Brainlesion Workshop. Springer, Cham, 2019. [121]

Authors: Gowtham Krishnan Murugesan, Sahil Nalawade, Chandan Ganesh, Ben Wagner, Fang F. Yu, Baowei Fei, Ph.D., Ananth J. Madhuranthakam, Joseph A. Maldjian

Author contributions: Gowtham Krishnan Murugesan and Sahil Nalawade designed the work, analyzed and interpreted the data, and wrote the paper. Dr. Joseph Maldjian, Dr. Ananth Madhuranthakam, Dr. Baowei Fei, and Dr. Fang F Yu, provided expert knowledge and mentorship to develop the method. Ben Wagner contributed in developing MRI analysis. Chandan Ganesh contributed to review the paper

5.1 Abstract.

In this work, we developed multiple 2D and 3D segmentation models with multiresolution input to segment brain tumor components and then ensembled them to obtain robust segmentation maps. Ensembling reduced overfitting and resulted in a more generalized model. Multiparametric MR images of 335 subjects from the BRATS 2019 challenge were used for training the models. Further, we tested a classical machine learning algorithm with features extracted from the segmentation maps to classify subject survival range. Preliminary results on the BRATS 2019 validation dataset demonstrated excellent performance with DICE scores of 0.898, 0.784, 0.779 for the whole tumor (WT), tumor core (TC), and enhancing tumor (ET), respectively and an accuracy of 34.5% for predicting survival. The Ensemble of multiresolution 2D networks achieved 88.75%, 83.28% and 79.34% dice for WT, TC, and ET respectively in a test dataset of 166 subjects

Keywords: Residual Inception Dense Networks, Densenet-169, Squeezenet, Survival Prediction, brain tumor segmentation

5.2 Introduction:

Brain Tumors account for 85-90% of all primary CNS tumors. The most common primary brain tumors are gliomas, which are further classified into a high grade (HGG) and low grade gliomas (LGG) based on their histologic features. Magnetic Resonance Imaging (MRI) is a widely used modality in the diagnosis and clinical treatment of gliomas. Despite being a standard imaging modality for tumor delineation and treatment planning, brain tumor segmentation on MR images remains a challenging task due to the high variation in tumor shape, size, location, and particularly the subtle intensity changes relative to the surrounding normal brain tissue. Consequently, manual tumor contouring is performed, which is both time-consuming and subject to large inter- and intra-observer variability. Semi- or fully-automated brain tumor segmentation methods could circumvent this variability for better patient management [122, 123]. As a result, developing automated, semi-automated, and interactive segmentation methods for brain tumors has important clinical implications, but remains highly challenging. Efficient deep learning algorithms to segment brain tumors into their subcomponents may help in early clinical diagnosis, treatment planning, and follow-up of patients [124].

The multimodal Brain Tumor Segmentation Benchmark (BRATS) dataset provided a comprehensive platform by outsourcing a unique brain tumor dataset with known ground truth segmentations performed manually by experts [8]. Several advanced deep learning algorithms were developed on this unique platform provided by BRATS and benchmarked against standard datasets allowing comparisons between them. Convolutional Neural Networks (CNN)-based methods have shown advantages for learning the hierarchy of complex features and have performed the best in recent BRATS challenges. U-net [125] based network architectures have been used for segmenting complex brain tumor structures. Pereira et al. developed a 2D CNN method with two CNN architectures for HGG and LGG separately and combined the outputs in the post-processing steps [126]. Havaei et al. developed a multi-resolution cascaded CNN architecture with two pathways, each of which takes different 2D patch sizes with four MR sequences as channels [127]. The

BRATS 2018 top performer developed a 3D decoder encoder style CNN architecture with inter-level skip connections to segment the tumor [128]. In addition to the decoder part, a Variation Autoencoder (VAE) was included to add reconstruction loss to the model.

In this study, we propose to ensemble output from Multiresolution and Multidimensional models to obtain robust tumor segmentations. We utilized off-the-shelf model architectures (DensNET-169, SERESNEXT-101, and SENet-154) to perform segmentation using 2D inputs. We also implemented a 2D and 3D Residual Inception Densenet (RID) network to perform tumor segmentation with patch-based inputs (64x64 and 64x64x64). The outputs from the model trained on different resolutions and dimensions were combined to eliminate false positives and post-processed using cluster analysis to obtain the final outputs.

5.3 Materials and Methods:

5.3.1 Data and Preprocessing

The BRATS 2019 dataset included a total of 335 multi-institutional subjects [111, 112, 129-131], consisting of 259 HGGs and 76 LGGs. The standard preprocessing steps by the BRATS organizers on all MR images included co-registration to an anatomical template [113], resampling to isotropic resolution ($1\times1\times1$ mm^3), and skull-stripping [114]. Additional preprocessing steps included N4 bias field correction [115] for removing RF inhomogeneity and normalizing the multi-parametric MR images to zero mean and unit variance.

The purpose of the survival prediction task is to predict the overall survival of the patient based on the multiparametric pre-operative MR imaging features in combination with the segmented tumor masks. Survival prediction based on only imaging-based features (with age and resection status) is a difficult task.

Additional information such as histopathology, genomic information, radiotracer based imaging, and other non-MR imaging features can be used to improve the overall survival prediction. Pooya et.. al. [132] reported better accuracy by combining genomic information and histopathological images to form a genomic survival convolutional neural network architecture (GSCNN model). Several studies have reported predicting overall survival for cerebral gliomas using ^{11}C-acetate and ^{18}F-FDG PET/CT scans [133-135].

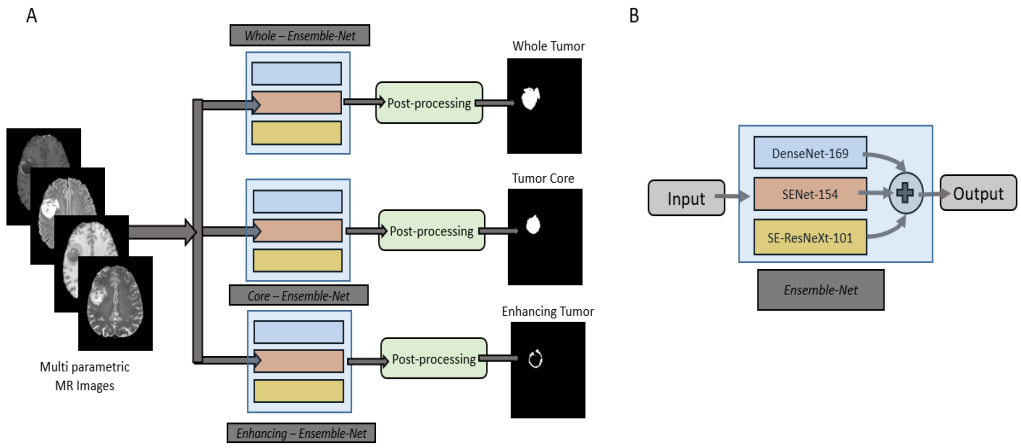


Figure 24: A. Ensemble of Segmentation models (DenseNET-169, SERESNEXT-101 and SENet-154). B. Ensemble methodology used to combine the outputs from Segmentation Models to produce output segmentation maps

5.3.2 Network Architecture

We trained several models to segment tumor components. All network architectures used for the segmentation task, except Residual Inception dense Network, were imported using Segmentation models, a python package [136]. The models selected for brain tumor segmentation had different backbones (DenseNet-169 [137], SERESNEXT-101 [138] and SENet-154 [139]). The DenseNet architecture has shown promising results in medical data classification and image segmentation tasks [140-142]. The DenseNet model has advantages in feature propagation from one dense block to the next and overcomes the problem of the vanishing gradient [137]. The squeeze and excitation block was designed to improve the feature propagation by enhancing the interdependencies between features for the classification task. This helps in propagating more useful features to the next block and suppressing less informative features. This network architecture was the top performer at the ILSVC 2017 classification challenge. SENet-154 and SE-

ResNeXt-101 have more parameters and is computationally expensive but has shown good results on the ImageNet classification tasks [139]. Three of the proposed models were ensembled to obtain the final results. All of these models from the Segmentation Models package were trained with 2D axial slices of size 240×240 (Fig. 24).

The Residual Inception Dense Network (RID) was first proposed and developed by Khened et al. for cardiac segmentation. We incorporated our implementation of the RID network in Keras with a Tensorflow backend (Figure 25). In the DenseNet architecture, the GPU memory footprint increases with the number of feature maps of larger spatial resolution. The skip connections from the down-sampling path to the up-sampling path use element-wise addition in this model, instead of the concatenation operation in DenseNet, to mitigate feature map explosion in the up-sampling path. For the skip connections, a projection operation was performed using Batch Normalization (BN)-1 × 1-convolution-dropout to match the dimensions for element-wise addition (Figure 26). These additions to the Densenet architecture help in reducing the parameters and the GPU memory footprint without affecting the quality of segmentation output. In addition to performing dimension reduction, the projection operation facilitates learning interactions of cross channel information [116] and faster convergence. Further, the initial layer of the RID networks includes parallel CNN branches similar to the inception module with multiple kernels of varying receptive fields. The inception module helps in capturing view-point dependent object variability and learning relations between image structures at multiple-scales.

5.3.3 Model Training and Ensemble Methodology.

All models from the Segmentation models package were trained with full resolution axial slices of size 240x240 as input to segment the tumor subcomponents separately. The outputs of each component from the models were combined following post-processing steps that included removing clusters of smaller size

to reduce false positives. Each tumor component was then combined to form the segmentation map (Figure 24B).

The RID model was trained on 2D input patches of size 64x64. For each component of the brain tumor (e.g., Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET)), we trained a separate RID model with axial as well as sagittal slices as input. In addition to the six RID models, we also trained a RID with axial slices as input with a patch size of 64x64 to segment TC and Edema simultaneously (TC-ED). A three-dimensional RID network model was also trained to segment ET and a multiclass TC-ED (TC-ED-3D). All models were trained with dice loss and Adam optimizers with a learning rate of 0.001 using NVIDIA Tesla P40 GPU's.

5.3.4 Ensemble Methodology.

The DenseNET-169, SERESNEXT-101, and SENet-154 model outputs were first combined to form segmentation maps, as shown in Figure 24B, which we will refer to as the Segmentation model output. Then, for each component, we combined outputs from the RID models and Segmentation models, as shown in Figure 26.

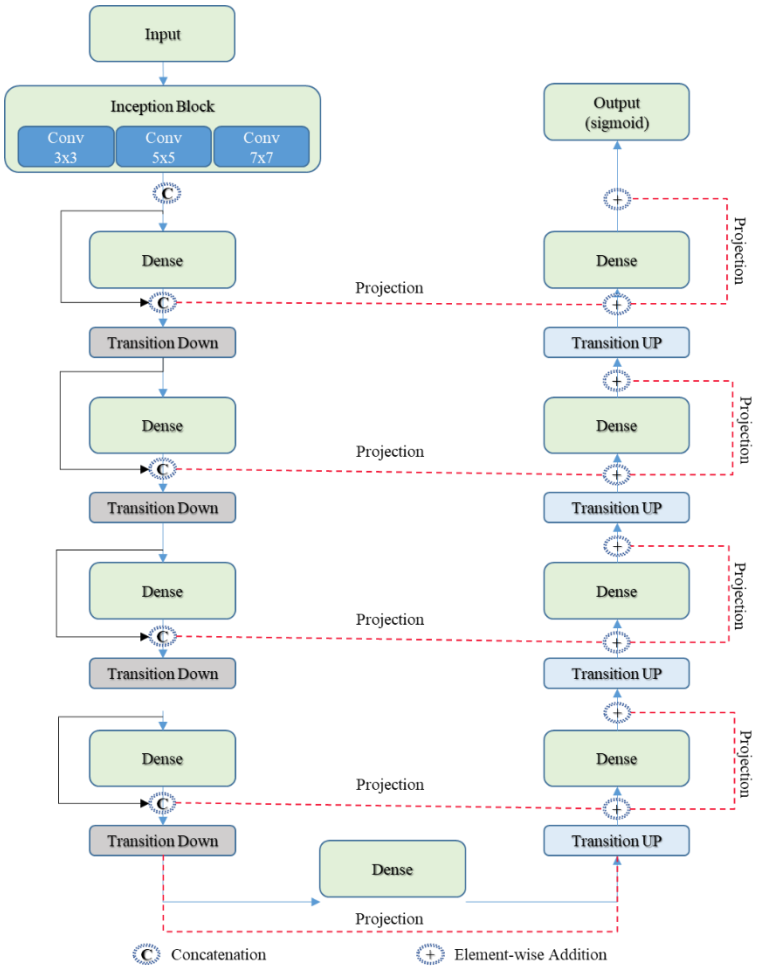


Figure 25: Residual Inception Densenet Architecture

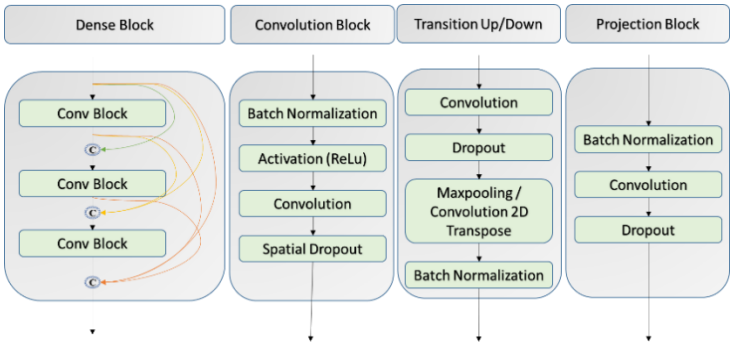


Figure 26: Building Blocks of Residual Inception Network. From left to right, dense block, convolution block, transition block and projection block

5.3.5 Survival Prediction.

The tumor segmentation maps extracted from the above methodology was used to extract texture and wavelet based features using the PyRadiomics [143] and Pywavelets [144] packages from each tumor subcomponent for each contrast. In addition, we also added volume and surface area features of each tumor component [145], along with age. We performed feature selection based on SelectKBest features using the sklearn package [146, 147], which resulted in a reduced set of 25 features. We trained four different models, including XGBoost (XGB), K-Nearest Neighbour (KNN), Extremely randomized trees (ET), and Linear Regression (LR) models [148] for the survival classification task. An ensemble of the four different models was used to form a voting classifier to predict survival in days. These predictions for each subject were then separated into low (<300 days), medium (300-450 days), and long survivors (>450 days). Twenty-nine subjects from the validation dataset were used to validate the trained model.

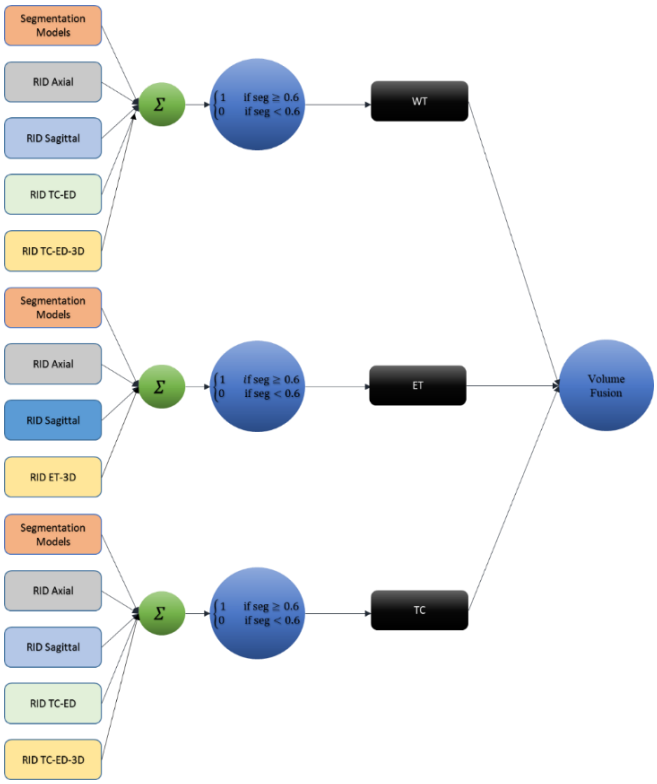


Figure 27: An ensemble of multidimensional and multiresolution networks. Top to bottom, the ensemble for the Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET), respectively.

5.4 Results

5.4.1 Segmentation

The Ensemble of multiresolution 2D networks achieved 89.79%, 78.43% and 77.97% dice for WT, TC, and ET respectively in the validation dataset of 125 subjects (Table 8, Fig. 27) and 88.75%, 83.28% and 79.34% dice for WT, TC, and ET respectively in the test dataset of 166 subjects (Table 9).

Table 8. Validation Segmentation Results for Multiresolution 2D ensemble model and multidimensional multiresolution ensemble model

Models	WT	TC	ET
Multiresolution 2D Ensemble	0.892	0.776	0.783
Multidimensional and Multiresolution Ensemble	0.898	0.78	0.784

Table 9. Testing Segmentation Results for the Multidimensional and Multiresolution ensemble model

Models	WT	TC	ET
Multidimensional and Multiresolution Ensemble	0.888	0.833	0.793

5.4.2 Survival Prediction

Accuracy and mean square error for overall survival prediction for the 29 subjects using a Voting Classifier were 51.7 % and 117923.1, respectively (Table 10). In testing, the proposed method achieved 41.1% accuracy.

Table 10. Validation Survival results for the Voting Classifier network.

	Accuracy	Mean Squared Error
Validation	51.7%	117923.1
Testing	41.1%	446765.3

5.4.3 Discussion

We ensemble several models with multiresolution inputs to segment brain tumors. The RID network was parameter and memory efficient, and able to converge in as few as three epochs. This allowed us to train several models for ensemble in a short amount of time. The proposed methodology of combining multidimensional models improved performance and achieved excellent segmentation results, as shown in Table 8. For survival prediction, we extracted numerous features based on texture, first-order statistics,

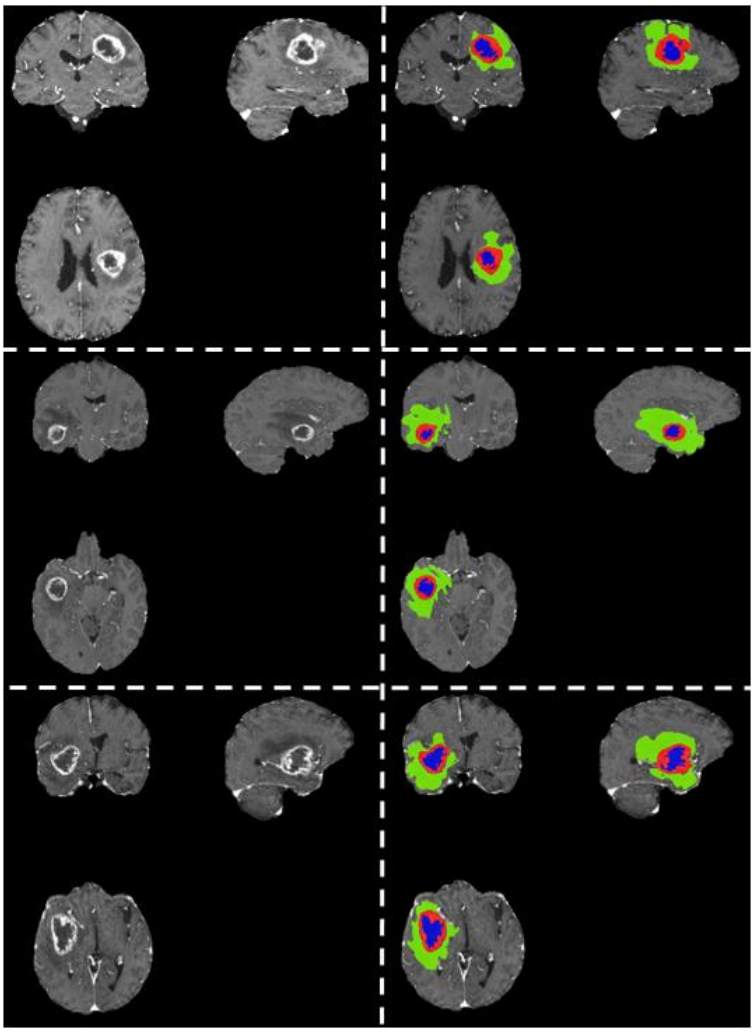


Figure 28:. Example Tumor Segmentation Performance for 3 subjects shown in each row. (a) T1-post contrast (T1C), (b) Segmentation output, (c) Overlay of segmentation output on the T1-post contrast images. Colors: Blue = Non-enhancing tumor + Necrosis, Red = Enhancing Tumor, and Green = Edema

and wavelets. Efficient model based feature selection allowed us to reduce the otherwise large feature set to 25 features per subject. We trained several classical machine learning models and then combined them to improve results on the validation dataset.

5.5 Conclusion

We demonstrated a two-dimensional multiresolution ensemble network for automated brain tumor segmentation to generate robust segmentation of tumor subcomponents. We also predicted the overall survival based on the segmented mask using an xgboost model. These may assist in diagnosis, treatment planning, and therapy response monitoring of brain tumor patients with more objective and reproducible measures.

5.6 References

1. Bakas, S., H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos (2017). Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017).
2. Bakas, S., H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos (2017). "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection." The Cancer Imaging Archive **286**.
3. Bakas, S., H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos (2017). "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features." Scientific data **4**: 170117.
4. Bakas, S., H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani and C. J. S. d. Davatzikos (2017). "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features." **4**: 170117.
5. Bakas, S., M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha and M. J. a. p. a. Rozycki (2018). "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge."
6. Buitinck, L., G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort and J. J. a. p. a. Grobler (2013). "API design for machine learning software: experiences from the scikit-learn project."
7. Chen, C.-F., Q. Fan, N. Mallinar, T. Sercu and R. J. a. p. a. Feris (2018). "Big-little net: An efficient multi-scale feature representation for visual and speech recognition."
8. Chen, L., Y. Wu, A. M. DSouza, A. Z. Abidin, A. Wismüller and C. Xu (2018). MRI tumor segmentation with densely connected 3D CNN. Medical Imaging 2018: Image Processing, International Society for Optics and Photonics.
9. Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM.
10. Dolz, J., K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers and I. B. J. I. t. o. m. i. Ayed (2018). "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation." **38**(5): 1116-1126.
11. Feng, X., N. Tustison and C. Meyer (2018). Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. International MICCAI Brainlesion Workshop, Springer.

12. Havaei, M., A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin and H. Larochelle (2017). "Brain tumor segmentation with deep neural networks." Medical image analysis **35**: 18-31.

13. Hu, J., L. Shen and G. Sun (2018). Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition.

14. Huang, G., Z. Liu, L. Van Der Maaten and K. Q. Weinberger (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition.

15. Islam, J. and Y. J. a. p. a. Zhang (2017). "An Ensemble of Deep Convolutional Neural Networks for Alzheimer's Disease Detection and Classification."

16. Kim, S., D. Kim, S. H. Kim, M.-a. Park, J. H. Chang, M. J. E. j. o. n. m. Yun and m. imaging (2018). "The roles of 11 C-acetate PET/CT in predicting tumor differentiation and survival in patients with cerebral glioma." **45**(6): 1012-1020.

17. Lee, G. R., R. Gommers, F. Waselewski, K. Wohlfahrt and A. J. J. O. S. S. O'Leary (2019). "PyWavelets: A Python package for wavelet analysis." **4**(36): 1237.

18. Lin, M., Q. Chen and S. Yan (2013). "Network in network." arXiv preprint arXiv:1312.4400.

19. Menze, B. H., A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom and R. Wiest (2014). "The multimodal brain tumor image segmentation benchmark (BRATS)." IEEE transactions on medical imaging **34**(10): 1993-2024.

20. Menze, B. H., A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom and R. J. I. t. o. m. i. Wiest (2014). "The multimodal brain tumor image segmentation benchmark (BRATS)." **34**(10): 1993-2024.

21. Mobadersany, P., S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. V. Vega, D. J. Brat and L. A. J. P. o. t. N. A. o. S. Cooper (2018). "Predicting cancer outcomes from histology and genomics using convolutional networks." **115**(13): E2970-E2979.

22. Myronenko, A. (2018). 3D MRI brain tumor segmentation using autoencoder regularization. International MICCAI Brainlesion Workshop, Springer.

23. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. J. J. o. m. l. r. Dubourg (2011). "Scikit-learn: Machine learning in Python." **12**(Oct): 2825-2830.

24. Pereira, S., A. Pinto, V. Alves and C. A. Silva (2016). "Brain tumor segmentation using convolutional neural networks in MRI images." IEEE transactions on medical imaging **35**(5): 1240-1251.

25. Rohlfing, T., N. M. Zahr, E. V. Sullivan and A. J. H. b. m. Pfefferbaum (2010). "The SRI24 multichannel atlas of normal adult human brain structure." **31**(5): 798-819.

26. Ronneberger, O., P. Fischer and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, Springer.

27. Saouli, R., M. Akil and R. Kachouri (2018). "Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images." Computer methods and programs in biomedicine **166**: 39-49.

28. Tsuchida, T., H. Takeuchi, H. Okazawa, T. Tsujikawa, Y. J. N. m. Fujibayashi and biology (2008). "Grading of brain glioma with 1-11C-acetate PET: comparison with 18F-FDG PET." **35**(2): 171-176.

29. Tustison, N. J., P. A. Cook, A. Klein, G. Song, S. R. Das, J. T. Duda, B. M. Kandel, N. van Strien, J. R. Stone and J. C. J. N. Gee (2014). "Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements." **99**: 166-179.

30. Van Griethuysen, J. J., A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper and H. J. J. C. r. Aerts (2017). "Computational radiomics system to decode the radiographic phenotype." **77**(21): e104-e107.

31. Yakubovskiy, P. (2019). "Segmentation Models." GitHub repository.

32. Yamamoto, Y., Y. Nishiyama, N. Kimura, R. Kameyama, N. Kawai, T. Hatakeyama, M. Kaji, M. J. M. I. Ohkawa and Biology (2008). "11 C-acetate PET in the evaluation of brain glioma: Comparison with 11 C-methionine and 18 F-FDG-PET." **10**(5): 281.

33. Zhuge, Y., A. V. Krauze, H. Ning, J. Y. Cheng, B. C. Arora, K. Camphausen and R. W. Miller (2017). "Brain tumor segmentation using holistically nested neural networks in MRI images." Medical physics **44**(10): 5234-5243.

6 Classification of Brain Tumor IDH Status using MRI and Deep Learning

Published: Nalawade, Sahil, et al. "Classification of brain tumor isocitrate dehydrogenase status using MRI and deep learning." Journal of Medical Imaging 6.4 (2019): 046003.[149]

Authors : Sahil Nalawade M.S., Gowtham Murugesan M.S., Maryam Vejdani-Jahromi, M.D., Ph.D., Ryan A. Fisicaro M.D., Chandan Ganesh Bangalore Yogananda, M.S., Ben Wagner B.M.E., Bruce Mickey, M.D., Elizabeth Maher, M.D., Ph.D., Marco C. Pinho, M.D., Baowei Fei, PhD, Ananth J. Madhuranthakam Ph.D., Joseph A. Maldjian, M.D.

Author contributions: Gowtham Krishnan Murugesan and Sahil Nalawade equally contributed and designed the work, analyzed and interpreted the data, and wrote the paper. Dr. Joseph Maldjian, Dr. Ananth Madhuranthakam, Dr. Elizabeth Maher, and Dr. Marco C. Pinho provided expert knowledge and mentorship to develop the method. Ben Wagner contributed in developing MRI analysis. Others, contributed to review the paper

6.1 Abstract

Isocitrate dehydrogenase (IDH) mutation status is an important marker in glioma diagnosis and therapy. We propose a novel automated pipeline for predicting IDH status noninvasively using deep learning and T2-weighted (T2w) MR images with minimal preprocessing (N4 bias correction and normalization to zero mean and unit variance). T2w MRI and genomic data were obtained from The Cancer Imaging Archive dataset (TCIA) for 260 subjects (120 High grade and 140 Low grade gliomas). A fully automated 2D densely connected model was trained to classify IDH mutation status on 208 subjects and tested on another held-out set of 52 subjects, using 5-fold cross validation. Data leakage was avoided by ensuring subject separation during the slice-wise randomization. Mean classification accuracy of 90.5% was achieved for each axial slice in predicting the three classes of no tumor, IDH mutated and IDH wild-type. Test accuracy of 83.8% was achieved in predicting IDH mutation status for individual subjects on the test dataset of 52 subjects. We demonstrate a deep learning method to predict IDH mutation status using T2w MRI alone. Radiologic imaging studies using deep learning methods must address data leakage (subject duplication) in the randomization process to avoid upward bias in the reported classification accuracy.

Keywords: Isocitrate dehydrogenase (IDH), MRI, Convolutional networks, Deep learning, Tumor classification, Radiomics

6.2 Introduction

In 2008 it was reported that some glioblastomas harbor a mutation in a gene coding for the citric acid cycle enzyme isocitrate dehydrogenase (IDH) [150]. Subsequent studies revealed that the majority of low grade gliomas possess a mutant form of IDH, and that the mutant enzyme catalyzes the production of the oncometabolite 2-hydroxyglutarate (2-HG) [151]. Although this product of the mutant form of IDH is believed to play a role in the initiation of the neoplastic process, it has been observed that gliomas that contain the mutant enzyme have a better prognosis than tumors of the same grade that contain only the wild type IDH. This observation implies that IDH mutated and IDH wild type gliomas are biologically different tumors, and led the World Health Organization (WHO) to designate them as such in the latest revision of their classification of gliomas [152]. Although a presumptive diagnosis of an IDH mutated glioma may be made on the basis of MR spectroscopy for 2-HG [153-156], at the present time, the only way to definitively identify an IDH mutated glioma is to perform immunohistochemistry or gene sequencing on a tissue specimen, acquired through biopsy or surgery. Because the differences between IDH mutated and IDH wild type gliomas may have implications for their treatment, especially if inhibitors of the mutant IDH enzyme currently in development prove to halt their growth, there is interest in attempting to distinguish between these two tumor types prior to surgery. As noted above, one avenue of research involves using MR spectroscopy to measure levels of 2-HG in the tumor [154, 157-159]. More recent studies have attempted to utilize machine learning techniques to analyze diagnostic MR images and predict IDH mutation status in gliomas using anatomic differences between the two tumor types.

Delfanti et al. demonstrated that genomic information with fluid attenuated inversion recovery (FLAIR) MRI could be used for the classification of patient images into IDH wild type, and IDH mutation with and without 1p/19q co-deletion [160]. The main determinants for classification were tumor border and location, with IDH mutant tumors having well-defined or slightly ill-defined borders and predominantly a frontal localization; and IDH wild type tumors demonstrating undefined borders and location in non-frontal areas. Chang et al. developed a deep learning residual network model for predicting IDH mutation with preprocessing steps including resampling, co-registration of multiple sequences, bias correction, normalization and tumor segmentation [161]. Using a combination of imaging and age, the model demonstrated testing accuracy of 89.1% and an area under the curve (AUC) value of 0.95 for IDH mutation for all image sequences combined. Zhang et al. used 103 low grade glioma (LGG) subjects for training a support vector machine (SVM) for classifying IDH mutation status, achieving an AUC of 0.83 on testing data [162]. In another approach, Chang et al [163] similarly demonstrated that IDH mutation status can be determined using T2-weighted (T2w), T2w- Fluid attenuated inversion recovery (FLAIR) and T1-weighted pre- and post-contrast images. Preprocessing steps in their work included co-registration of all sequences, intensity normalization using zero mean and unit variance, application of a 3D convolutional neural network (CNN) based whole tumor segmentation tool for segmenting the lesion margins, cropping the output tumor mask on all input imaging sequences, and resizing individual image slices to 32 x 32 with 4 input sequence channels. The mean accuracy result from the model was 94% with a 5-fold cross validation accuracy ranging from 90% to 96% [163]. Common to all of these previous methods is the involvement of preprocessing steps, typically including some form of brain tumor pre-segmentation or region of interest extraction,

and utilizing multiparametric or 3D near-isotropic MRI data that is often not part of the standard clinical imaging protocol [161, 163].

In this work, we propose a fully automated deep learning based pipeline using a densely connected network model, that involves minimal preprocessing and requires only standard T2w images. A similar approach has been previously used for the identification of the O^6 – methylguanine-DNA methyltransferase (MGMT) methylation status and prediction of 1p/19q chromosomal arm deletion [164]. Clinical T2-weighted images are acquired in a short time frame (typically around 2 minutes), and are robust to motion with current acquisition methods. Almost universally, high quality T2-weighted images are acquired during clinical brain tumor work-ups. The preprocessing steps preserve the original image information without the need for any resampling, skull stripping, region-of-interest, or tumor pre-segmentation procedures. The advantage of a dense network model is that it passes the weights from all the previous blocks to the subsequent blocks, preserving the information from the initial layer and aiding in the classification.

The ability to quickly and accurately classify IDH status non-invasively can help with better planning, counseling, and treatment of brain tumor patients, especially in cases where biopsy is not feasible due to unfavorable tumor locations. A methodologic contribution that we make specifically to the radiologic deep learning literature is on the approach to data randomization for 2D models. Furthermore, the deep learning approach is fully automated and can be easily implemented in the clinical workflow using only T2-weighted MR images.

6.3 Materials and Methods

6.3.1 Subjects

260 subjects from The Cancer Imaging Archive (TCIA) [165] dataset were selected, including 120 high grade gliomas (HGG) [166] and 140 low grade gliomas (LGG) [167], and based on their pre-operative status from a pool of 461 subjects. The genomic information was provided through the National Cancer Institute - Genomic Data Commons (GDC) Data Portal [168]. The genomic data was available in the following 3 classes: IDH mutated, IDH wild type, and Not Available (N/A). The Genomic data of the N/A type was excluded from the pool of 461 subjects. MRI data was filtered for any visible artifacts in the images. The final dataset consisted of 260 subjects based on the available genomic information, MRI data, pre-operative status and lack of image artifacts on the T2w images.

A standard 80:20 data split was employed with 80% training and 20% testing (held-out). The 80% training was further split into a standard 80:20 split of 80% training and 20% validation. The final dataset of 260 subjects was thus randomly divided into a training set (208 subjects, including approximately 96 HGG and 112 LGG) and a test set (52 subjects, including approximately 24 HGG and 28 LGG). This process was repeated separately for each fold during the 5-fold cross validation.

For each fold of the cross-validation, 208 subjects with, on average, 9,728 axial slices of T2w images were selected for training and validation (7177 slices – No tumor, 1110 slices – IDH mutated, 1441 slices– IDH wild type). The start and end slices of the tumor (edge slices) were manually labeled for each T2 dataset. These edge slices were excluded from training to provide more robust ground truth data. All slices were included for the testing set. Each T2w slice was

manually assigned only one label (No Tumor, IDH mutated, or IDH wild type). In order to address any class imbalance due to the higher number of no tumor slices, class weights were assigned based on the labels in the training dataset. Although this was a slice-wise training model, slices of subjects in the testing set were not mixed into the training set. This is a critical step related to the data leakage problem in 2D networks, especially for radiologic deep learning studies [169, 170]. This was necessary to avoid bias during testing and an over inflation of the measured accuracies. Fifty two subjects with 2522 axial slices (1839 slices– No Tumor, 299 slices– IDH mutated, 384 slices– IDH wild type) were not included in the training or validation and were used for testing, for each fold. Classification was done on a slice-wise basis (2D) followed by majority voting across all slices to provide a patient-level classification. Note that we use the term slice-wise to refer to classification of each 2D axial image for IDH status. Similarly, the term subject-wise is used for classification of IDH status for each subject. We used a straightforward majority voting scheme to determine subject-wise classification based on the majority IDH classification of the individual 2D slices. Subjects classified with an equal number of IDH mutated and IDH wild-type tumor slices were assigned to the IDH wild type group.

6.3.2 Image Processing

Minimal standard preprocessing of the T2w images from the TCIA data set was performed prior to training (Figure 29). The images were converted from DICOM to nifti format using dcm2nii, bias corrected to remove RF inhomogeneity using the N4 bias correction algorithm, zero-mean intensity normalized to between -1 and 1, and resampled to 128 x 128 image dimensions to improve the computational efficiency during training. The Inception V4 model however, required

input image size of 299 x 299 as a design constraint of this model when originally constructed [171, 172]. The total preprocessing time for each subject was less than 1 minute.

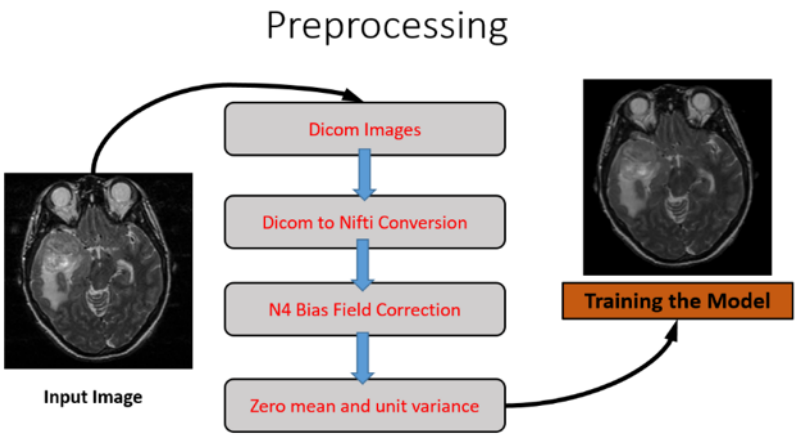


Figure 29: Flowchart of preprocessing steps prior to training the deep learning

6.3.3 Model Training

The following models were used for classification of the T2w images into IDH mutated and IDH wild type classes: residual network (ResNet-50), Densely connected network (DenseNet-161), and Inception-v4. Our choice of network architectures was based on the best performers from the ImageNet challenge for 2015 (ResNET), and 2017 (DenseNet and Inception V4). The DenseNet model, designed by Huang et al [173] received the best paper award at CVPR 2017. The models were trained with the Pycharm and Python IDEs using the Keras python package with TensorFlow backend engines. Fine tuning of the 3 classes was performed on all models. The three-class labels for each slice were: no tumor, IDH mutated, and IDH wild type. The models were originally trained on ImageNet data with 3 channels (RGB). For our implementation the 3-channel input was provided as a central slice with the 2 immediate surrounding slices. If the central slice was the first or last slice, the surrounding slices were assigned as no value.

6.3.4 ResNet-50 Model

The residual network was implemented as proposed by He et al. [174]. Each residual connection adds the input of the block to the output, helping to preserve the information from the previous block. A deep residual network framework was added to the model while maintaining parameter numbers to address issues with convergence in the originally proposed model. The residual net used the kernel initializer as 'He normal' for weight initialization. On top of the residual network model, a flattened output was added and sent to the dense layer with the rectified linear unit ('relu') activation and a dropout of 0.5. The final layer of the model was the classification layer with a softmax activation and the number of classes as the output. The residual network model used for training was ResNet-50 (Figure 30).

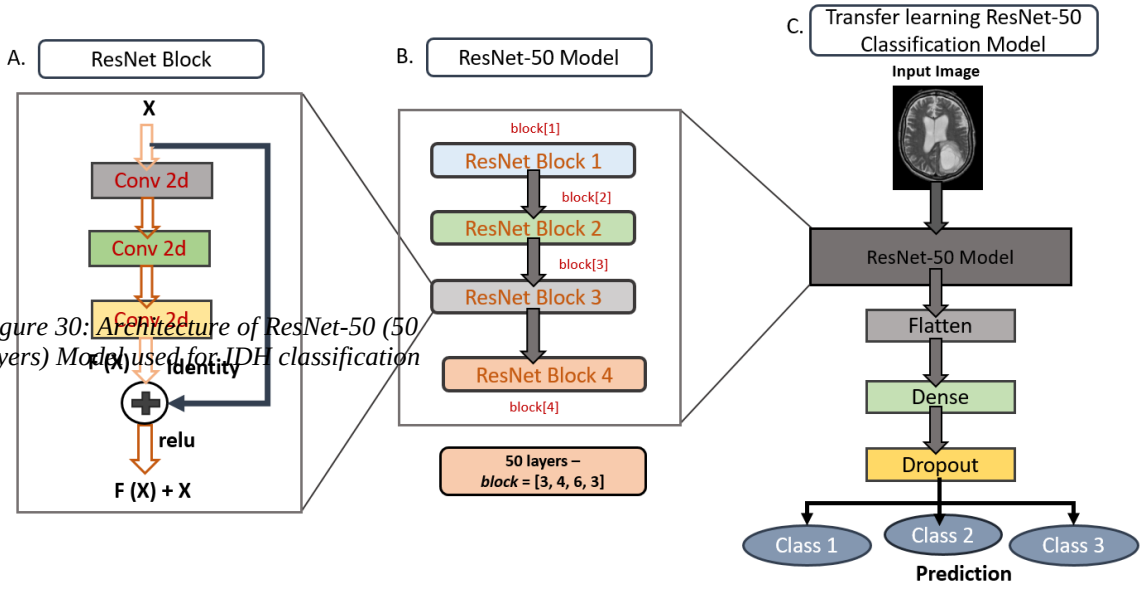


Figure 30: Architecture of ResNet-50 (50 layers) Model used for IDH classification

6.3.5 Inception-v4 Model

The Inception model architecture was designed by the Google Research team [171, 172]. The Inception-v4 model is a deep architecture with 41 million parameters and the model is designed

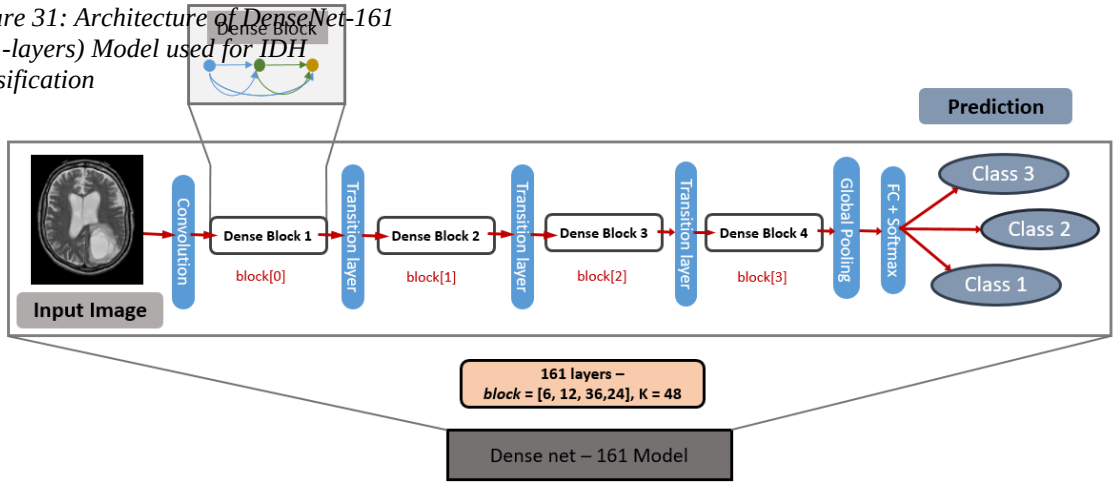
with inception blocks and reduction blocks. The inception blocks are used in a sequential manner with reduction blocks except for the last inception block, which has an average pooling layer and a dropout layer before the classification layer.

6.3.6 DenseNet-161 Model

The DenseNet model was based on the design by Huang et al. [173]. This model was inspired by the residual network model, which allows the residual connections to pass information from the previous layer to the subsequent layer. Dense networks have advantages over other networks by alleviating the vanishing gradient problem with feature propagation through the dense connection to the subsequent layers.

The features passed to the subsequent layers in the DenseNet model are not added by summation, but are combined using concatenation. Each block has connections from the previous block such that L=number of blocks and the number of connections for each block is $L\times(L+1)/2$, creating a dense connectivity pattern or DenseNet. The DenseNet-161 model architecture is shown in Figure 31, which illustrates a 5 block approach where the 1st block is the Input layer and each of the subsequent 4 blocks are characterized by 2D convolution layers with filter size of (1 x 1) and (3 x 3) respectively. The pre-trained model was used to transfer learning and used for classification based on the trained information. A 161 layer DenseNet model was used for model training.

Figure 31: Architecture of DenseNet-161 (161-layers) Model used for IDH classification



6.3.7 Training, Testing and Statistical Analysis

Model training was performed on a Nvidia Tesla P100, P40, K40/K80 GPU with 384 GB RAM and the model accuracy was assessed for 200 Epochs. The optimizer used for training was the Stochastic Gradient Descent [175] as described in Zhang et al.[176] and the learning rate was set to 10^{-5}, with a decay of 10^{-7} and momentum of 0.8. Data augmentation was performed on the training dataset, which included vertical and horizontal flip, random rotation, translation, shear, zoom shifts and elastic transformation to minimize overfitting the data. The results were analyzed by assessing accuracy, precision, sensitivity, specificity, and F-1 score values. Figure 32 shows the confusion matrix and the equations for calculating the testing parameters. Slice-wise model testing was performed based on the output from the 2D model. Subject-wise classification was performed based on majority voting across IDH mutated and IDH wild type tumor slices. This classification accuracy was computed on the independent test dataset that was separate from the testing and validation data sets.

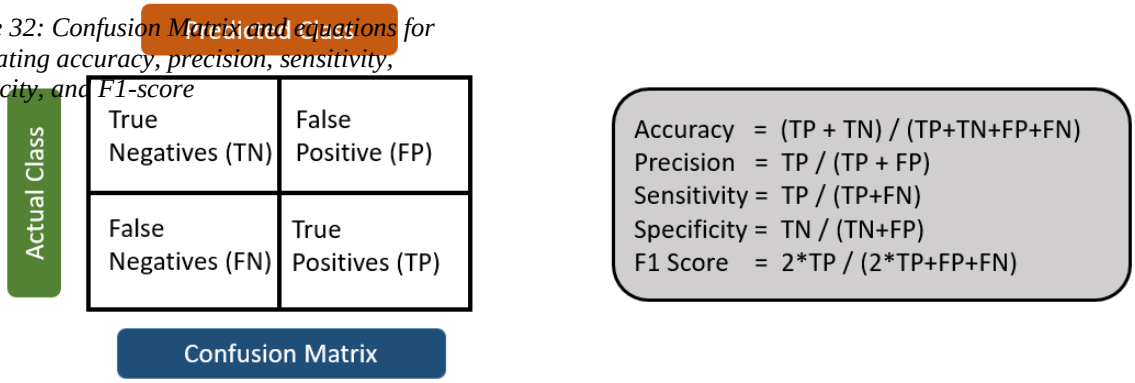


Figure 32: Confusion Matrix and equations for calculating accuracy, precision, sensitivity, specificity, and F1-score

6.3.8 Model training times

The DensetNet-161 model took approximately 110 hours for training, while the ResNet-50 model and the Inception V4 model took approximately 56 hours and 32 hours, respectively. Testing time for individual subject classification was less than 30 seconds for all models.

6.4 Results

6.4.1 Training, validation, and testing accuracy

Table 11 shows the accuracy comparison between the ResNet-50, DenseNet-161 and Inception-v4 models. The DenseNet-161 model was superior in training, validation, and testing accuracy compared to ResNet-50 and Inception-v4. Averaged across the five folds, the slice-wise accuracy of the DenseNet-161 model was 90.5 ±1.0% (standard deviation) with an AUC of 0.95 on the held out test dataset of 52 subjects.

Table 11: Slice-wise accuracy comparisons between the Resnet-50, Inception-v4, and DenseNet-161 model averaged for 5 fold cross validation

Results averaged for 5 fold cross validation			
Model	**Training accuracy (%)**	**Validation accuracy (%)**	**Testing accuracy (%)**
Inception-v4	64.8	72.2	1916/2522 (76.1)
ResNet-50	97.9	96.5	2265/2522 (89.7)
DenseNet-161	**97.9**	**96.4**	**2282/2522 (90.5)**

6.4.2 Accuracy, Precision, Recall/Sensitivity, Specificity, F1 score and AUC Comparison

Average metrics were computed across folds and classes. The classification accuracy, precision, recall/sensitivity, specificity, F1 score and AUC for slice-wise IDH classification with the DenseNet-161 model were 90.5 ±1.0%, 79.9 ±3.4%, 83.1 ±3.2%, 94.8 ±0.5%, 81.3 ±3.2% and 0.95, respectively. For subject-wise IDH classification, accuracy, precision/positive predictive value, recall/sensitivity, specificity, F1 score and AUC were 84.1 ±2.9%, 83.5 ±3.5%, 83.5 ±3.5%, 83.5 ±3.1%, and 0.84 (Table 12). Slice-wise and subject-wise comparisons of accuracy, precision, recall/sensitivity, specificity, F1 score and AUC for each of the 5 fold cross validations for the DenseNet-161 model are shown in Table 13.

Table 12: Slice-wise and subject-wise comparison of accuracy, precision, recall, F1-score and AUC parameters between the Resnet-50, Inception-v4, and DenseNet-161 model averaged for 5 fold cross validation

Results averaged for 5 fold cross validation						
Parameters	**Accuracy (%)**	**Precision (%)**	**Recall / Sensitivity (%)**	**Specificity (%)**	**F1 score (%)**	**AUC**
Slice wise						
Inception-v4	76.1	59.4	59.2	84.5	58.2	0.86
ResNet-50	89.7	79.3	81.7	94.1	80.2	0.95
DenseNet-161	**90.5**	**79.9**	**83.1**	**94.8**	**81.3**	**0.95**
Subject wise						
Inception-v4	64.2	65.8	65.1	65.1	64.0	0.65
ResNet-50	81.4	81.5	81.5	81.5	81.4	0.81
DenseNet-161	**83.8**	**84.1**	**83.5**	**83.5**	**83.5**	**0.84**

Table 13: Slice-wise and subject-wise comparison of accuracy, precision, recall, F1 score and AUC parameters for each of the fivefold cross validation for the DenseNet-161 model

DenseNet-161 model						
Fold	**Accuracy (%)**	**Precision (%)**	**Recall / Sensitivity (%)**	**Specificity (%)**	**F1 score (%)**	**AUC**
Slice wise						
1	91.7	83.3	86.0	94.9	84.4	0.95
2	91.0	82.7	84.4	95.1	83.5	0.95
3	90.1	79.3	81.5	94.5	80.2	0.95
4	88.7	73.7	77.6	93.9	75.5	0.91
5	90.9	80.3	86.0	95.4	82.8	0.95
Subject wise						
1	84.6	84.8	84.0	84.0	84.2	0.84
2	86.5	87.2	87.5	87.5	86.5	0.87
3	78.8	79.1	77.9	77.9	78.2	0.78
4	82.7	83.1	81.8	81.8	82.2	0.82
5	86.5	86.3	86.6	86.6	86.4	0.87

6.4.3 Slice-wise comparison:

The precision for the DenseNet-161 model across 5 fold cross validation was 97.7 ±0.5% for the "no tumor" classification, 71.7 ±6.8% for IDH mutation, and 70.3 ±5.5% for IDH wild type.

6.4.4 Subject-wise comparison:

For the DenseNet-161 model, the sensitivity and specificity for subject-wise IDH mutation classification was 80.9 ±9.4% and 86.2 ±3.8%, respectively. The positive and negative predictive values were 82.5 ±2.8% and 85.7 ±6.3%, respectively.

6.5 Discussion

The results from Tables 11 and 12 show that the ResNet-50 model performed better than the Inception-v4 model. The ResNet-50 architecture has residual connections which preserve information from the previous layer in the residual block. The DenseNet-161 model performed the best of all the three models tested. Unlike the ResNet-50 model, the DenseNet-161 model architecture carries the information from all previous layers and adds the information to the next layer. This helped in learning the information from different layers and transferring to the next layers. The slice-wise classification AUC results were 0.95 for DenseNet-161, 0.95 for ResNet-50, and 0.86 for Inception-v4.

Chang et al. [163] demonstrated a high classification accuracy for IDH mutation status using T2w, FLAIR, T1w pre- and post-contrast images. Preprocessing steps included coregistration across multiple sequences, intensity normalization to zero mean and unit variance, segmentation of the brain tumor and cropping the images and resizing slices to 32 x 32. A 94% mean accuracy on 5-

fold cross validation was reported. The approach to classification was slice-wise, similar to our model. In designing the slice-wise classification model, it is important to ensure that none of the slices of subjects from the testing set are inadvertently included in the training set. This can easily be overlooked in 2D slice-wise models during the slice randomization process that generate the training slices, validation slices, and testing slices. This can introduce bias in the testing phase artificially boosting accuracies by including slices from subjects in the training set that share considerable information with different slices but from the same subjects in the testing set. It is not clear in the previously reported 2D models whether this caveat was adhered to.

An important methodologic contribution that we make specifically to the radiologic deep learning literature is on the approach to data randomization for 2D models. It is critical that imaging researchers are aware of the data leakage and subject duplication issue. This is perhaps unique to radiology where multiple slices of pathology are acquired in MRI or CT, with considerable overlap in feature content from slice to slice. Widely used deep learning tools provide the ability to perform data randomization using a simple flag in the called routine (*e.g.*, in Keras, or Scikit-learn[146]). Use of this flag in 2D imaging based CNNs can lead to bias in the results by inadvertently including slices from the same subject in both training and testing cohorts. This is a significant concern, as it can lead to data leakage in which examples of the same subject (albeit different slices of the same tumor) can appear in the training set and the test set. The problem of data leakage in medical images was discussed by Wegmayr et. al. [177] and Feng et. al.[169] and has been referred to as subject duplication in training and testing sets. In our initial studies, we did not account for the data leakage problem and achieved accuracies of 95% with the T2 images alone, slightly higher than that of Chang et al [163]. When appropriately accounting for the data leakage issue, our accuracies were reduced to the 83.8% reported here. One of the major

contributions of our work is in making the radiology community aware of the data leakage problem, as it is very easy to overlook when 2D networks are considered that use image slices as input.

The majority of HGG tumors are IDH wild type (up to 90%). An algorithm that merely distinguishes between HGG and LGG for determination of IDH status is likely of limited value as this can be done subjectively with fairly high accuracy on the basis of contrast enhancement. For example, previous studies that used multiparametric MR data for **determination of IDH status in** HGG and LGG may have demonstrated high accuracy predominantly on the basis of contrast enhancement features. The more valuable distinction from a clinical standpoint would be between IDH mutated and IDH wild type low grade gliomas in which contrast enhancement is **usually absent.** Our training and testing **samples were** weighted towards LGG, and there were a significant number of **IDH wild type** LGG in both the training and validation sample (~ 30%). Our testing accuracy for the LGG group was 78.6%. Additionally, our use of T2w-only images eliminates the potential for the algorithm being a contrast-enhancement discriminator.

Our method provides high accuracy with minimal preprocessing steps as compared to previous work. The preprocessing steps in our work only involve N4 bias field correction and intensity normalization. Our method also involves no tumor segmentation or ROI extraction as described in Chang et al.[161] , which helps in reducing the time, effort and potential sources of error. Our method also does not require pre-engineered features to be extracted from the images or histopathological data as described in Delfanti et al. [160]. This general approach can be easily incorporated into an automated clinical workflow for IDH classification. The minimal

preprocessing, and the use of standard T2w images alone makes it promising as a robust clinical tool for noninvasively determining IDH mutation status.

Limitations

This is a retrospective study applying several neural network architectures to the TCIA HGG-LGG database to generate a model predicting IDH genotype based only on T2-weighted MR imaging. The data set, especially at the subject level, is small in terms of deep learning applications and may not generalize well. Fluctuation of performance is also a concern with small data sets. However, the TCIA dataset is the largest curated brain tumor dataset publicly available, and it uses data from multiple sites using different imaging protocols. This database consisted of data from 10 different institutions out of which 8 institutions contributed GBM/HGG datasets and 5 institutions contributed LGG datasets to the TCIA cohort. This provided a very heterogeneous dataset, and we believe this is perhaps even better than using data from a single source for deep learning applications. While our current study focused on the classification of T2w images into no tumor, IDH mutated, an IDH wild type, future studies can extend this approach to classify IDH1 and IDH2 subtypes. Accuracies may be further improved with the inclusion of multiparametric imaging data in the training model. Our approach, however, is much more straightforward using T2-weighted images alone without the requirement of additional imaging sequences. Clinically, T2-weighted images are typically acquired within 2 minutes, and are robust to patient motion. The multi-sequence input required by previous approaches can be compromised due to patient motion from lengthier examination times, and the need for gadolinium contrast, especially as the post-contrast images are typically acquired at the end of an already lengthy examination time. For a potential

clinical solution, the use of T2-weighted images is a significant strength, as these images are almost uniformly acquired without artifacts from patient motion.

6.6 Conclusion

We demonstrate a deep learning method to predict IDH mutation status using T2-weighted MR images alone. The proposed model requires minimal preprocessing to obtain high accuracies, without the need for tumor segmentation or extraction of regions of interest, making it promising for robust clinical implementation.

Disclosures

No conflicts of interest

Acknowledgments

Support for this research was provided by NCI U01CA207091 (AJM, JAM).

6.7 References

[1] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505-515, 2017.

[2] M. Kim *et al.*, "Deep Learning in Medical Imaging," *Neurospine*, vol. 17, no. 2, pp. 471-472, Jun 2020.

[3] J. Sui, M. Liu, J. H. Lee, J. Zhang, and V. Calhoun, "Deep learning methods and applications in neuroimaging," *J Neurosci Methods*, vol. 339, p. 108718, Jun 1 2020.

[4] J. W. Song, N. R. Yoon, S. M. Jang, G. Y. Lee, and B. N. Kim, "Neuroimaging-Based Deep Learning in Autism Spectrum Disorder and Attention-Deficit/Hyperactivity Disorder," *Soa Chongsonyon Chongsin Uihak*, vol. 31, no. 3, pp. 97-104, Jul 1 2020.

[5] A. Abrol *et al.*, "Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease," *J Neurosci Methods*, vol. 339, p. 108701, Jun 1 2020.

[6] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong, "Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review," *Comput Methods Programs Biomed*, vol. 187, p. 105242, Apr 2020.

[7] C. R. Jack Jr *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685-691, 2008.

[8] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993-2024, 2014.

[9] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498-1507, 2007.

[10] G. Zhu, B. Jiang, L. Tong, Y. Xie, G. Zaharchuk, and M. Wintermark, "Applications of deep learning to neuro-imaging techniques," *Frontiers in Neurology*, vol. 10, p. 869, 2019.

[11] J. Jurek, M. Kociński, A. Materka, M. Elgalal, and A. Majos, "CNN-based superresolution reconstruction of 3D MR images using thick-slice scans," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 111-125, 2020.

[12] M. J. Sheller *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific RepoRtS*, vol. 10, no. 1, pp. 1-12, 2020.

[13] F. Pesapane, C. Volonté, M. Codari, and F. Sardanelli, "Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States," *Insights into imaging*, vol. 9, no. 5, pp. 745-753, 2018.

[14] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31-57, 2018.

[15] O. Sporns, "Graph theory methods: applications in brain networks," *Dialogues in Clinical Neuroscience*, vol. 20, no. 2, p. 111, 2018.

[16] A. Fornito, A. Zalesky, and M. Breakspear, "The connectomics of brain disorders," *Nat Rev Neurosci*, vol. 16, no. 3, pp. 159-72, Mar 2015.

[17] A. Avena-Koenigsberger, B. Misic, and O. Sporns, "Communication dynamics in complex brain networks," *Nature Reviews Neuroscience*, vol. 19, no. 1, p. 17, 2018.

[18] R. L. Buckner *et al.*, "Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease," *Journal of neuroscience*, vol. 29, no. 6, pp. 1860-1873, 2009.

[19] N. A. Crossley *et al.*, "The hubs of the human connectome are generally implicated in the anatomy of brain disorders," *Brain*, vol. 137, no. 8, pp. 2382-2395, 2014.

[20] D. E. Warren *et al.*, "Network measures predict neuropsychological outcome after brain injury," *Proceedings of the National Academy of Sciences*, vol. 111, no. 39, pp. 14247-14252, 2014.

[21] A. Fornito, A. Zalesky, and E. Bullmore, *Fundamentals of brain network analysis*. Academic Press, 2016.

[22] S. M. Smith *et al.*, "Network modelling methods for FMRI," *Neuroimage*, vol. 54, no. 2, pp. 875-91, Jan 15 2011.

[23] W. H. Kim *et al.*, "Cerebrospinal fluid biomarkers of neurofibrillary tangles and synaptic dysfunction are associated with longitudinal decline in white matter connectivity: A multi-resolution graph analysis," *NeuroImage: Clinical*, vol. 21, p. 101586, 2019.

[24] W. H. Kim *et al.*, "Multi-resolution statistical analysis of brain connectivity graphs in preclinical Alzheimer's disease," *NeuroImage*, vol. 118, pp. 103-117, 2015.

[25] G. Zaharchuk, E. Gong, M. Wintermark, D. Rubin, and C. Langlotz, "Deep learning in neuroradiology," *American Journal of Neuroradiology*, vol. 39, no. 10, pp. 1776-1784, 2018.

[26] G. Murugesan *et al.*, "Single Season Changes in Resting State Network Power and the Connectivity between Regions: Distinguish Head Impact Exposure Level in High School and Youth Football Players," *Proc SPIE Int Soc Opt Eng*, vol. 10575, Feb 2018.

[27] Z. Zhou *et al.*, "A toolbox for brain network construction and classification (BrainNetClass)," 2020.

[28] L. Waller, A. Brovkin, L. Dorfschmidt, D. Bzdok, H. Walter, and J. D. J. J. o. n. m. Kruschwitz, "GraphVar 2.0: A user-friendly toolbox for machine learning on functional connectivity measures," vol. 308, pp. 21-33, 2018.

[29] N. Williams and R. N. Henson, "Recent advances in functional neuroimaging analysis for cognitive neuroscience," ed: SAGE Publications Sage UK: London, England, 2018.

[30] E. Pellegrini *et al.*, "Machine learning of neuroimaging to diagnose cognitive impairment and dementia: a systematic review and comparative analysis," *arXiv preprint arXiv:1804.01961*, 2018.

[31] T. J. O'Neill, E. M. Davenport, G. Murugesan, A. Montillo, and J. A. Maldjian, "Applications of resting state functional mr imaging to traumatic brain injury," *Neuroimaging Clinics*, vol. 27, no. 4, pp. 685-696, 2017.

[32] T. Turki, J. T. Wang, and I. Rajikhan, "Inferring gene regulatory networks by combining supervised and unsupervised methods," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 140-145: IEEE.

[33] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins, "Next-generation machine learning for biological networks," *Cell*, 2018.

[34] J. D. Finkle, J. J. Wu, and N. Bagheri, "Windowed Granger causal inference strategy improves discovery of gene regulatory networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 9, pp. 2252-2257, 2018.

[35] A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PloS one*, vol. 5, no. 9, p. e12776, 2010.

[36] W. Yan, H. Zhang, J. Sui, and D. Shen, "Deep chronnectome learning via full bidirectional long short-term memory networks for MCI diagnosis," in *International conference on medical image computing and computer-assisted intervention*, 2018, pp. 249-257: Springer.
[37] X. Chen *et al.*, "High-order resting-state functional connectivity network for MCI classification," vol. 37, no. 9, pp. 3282-3296, 2016.
[38] R. Yu, H. Zhang, L. An, X. Chen, Z. Wei, and D. J. H. b. m. Shen, "Connectivity strength-weighted sparse group representation-based brain network construction for M CI classification," vol. 38, no. 5, pp. 2370-2383, 2017.
[39] K. Hilger and C. J. Fiebach, "ADHD symptoms are associated with the modular structure of intrinsic brain networks in a representative sample of healthy adults," *Network Neuroscience*, vol. 3, no. 2, pp. 567-588, 2019.
[40] P. Lin *et al.*, "Global and local brain network reorganization in attention-deficit/hyperactivity disorder," *Brain imaging and behavior*, vol. 8, no. 4, pp. 558-569, 2014.
[41] F. Saeed, "Towards quantifying psychiatric diagnosis using machine learning algorithms and big fMRI data," *Big Data Analytics*, vol. 3, no. 1, p. 7, 2018.
[42] S. Cortese *et al.*, "Toward systems neuroscience of ADHD: a meta-analysis of 55 fMRI studies," *American Journal of Psychiatry*, vol. 169, no. 10, pp. 1038-1055, 2012.
[43] J. Sidlauskaite, K. Caeyenberghs, E. Sonuga-Barke, H. Roeyers, and J. R. Wiersema, "Whole-brain structural topology in adult attention-deficit/hyperactivity disorder: Preserved global–disturbed local network organization," *NeuroImage: Clinical*, vol. 9, pp. 506-512, 2015.
[44] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock, "The neuro bureau ADHD-200 preprocessed repository," *Neuroimage*, vol. 144, pp. 275-286, 2017.
[45] M. P. Milham, D. Fair, M. Mennes, and S. H. Mostofsky, "The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience," *Frontiers in systems neuroscience*, vol. 6, p. 62, 2012.
[46] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273-289, 2002.
[47] A. Abraham *et al.*, "Machine learning for neuroimaging with scikit-learn," *Frontiers in neuroinformatics*, vol. 8, p. 14, 2014.
[48] C. J. Stam and J. C. Reijneveld, "Graph theoretical analysis of complex networks in the brain," *Nonlinear biomedical physics*, vol. 1, no. 1, p. 3, 2007.
[49] L. Breiman, *Classification and regression trees*. Routledge, 2017.
[50] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics*, vol. 31, no. 12, pp. i197-i205, 2015.
[51] A. Yamashita *et al.*, "Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias," vol. 17, no. 4, p. e3000042, 2019.
[52] J.-P. Fortin *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," vol. 167, pp. 104-120, 2018.
[53] J. Wang, X. Wang, M. Xia, X. Liao, A. Evans, and Y. J. F. i. h. n. He, "GRETNA: a graph theoretical network analysis toolbox for imaging connectomics," vol. 9, p. 386, 2015.
[54] A. Hagberg *et al.*, "Networkx. High productivity software for complex networks," *Webová strá nka https://networkx. lanl. gov/wiki*, 2013.
[55] J. Wang, X. Zuo, and Y. J. F. i. s. n. He, "Graph-based network analysis of resting-state functional MRI," vol. 4, p. 16, 2010.
[56] S. Achard and E. J. P. c. b. Bullmore, "Efficiency and cost of economical brain functional networks," vol. 3, no. 2, p. e17, 2007.
[57] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3-42, 2006.
[58] M. Wehenkel, C. Bastin, C. Phillips, and P. Geurts, "Tree ensemble methods and parcelling to identify brain areas related to Alzheimer's disease," in *2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2017, pp. 1-4: IEEE.
[59] K. Supekar, M. Musen, and V. Menon, "Development of large-scale functional brain networks in children," *PLoS biology*, vol. 7, no. 7, p. e1000157, 2009.
[60] S. Achard and E. Bullmore, "Efficiency and cost of economical brain functional networks," *PLoS computational biology*, vol. 3, no. 2, p. e17, 2007.
[61] X. Qian *et al.*, "Large-scale brain functional network topology disruptions underlie symptom heterogeneity in children with attention-deficit/hyperactivity disorder," vol. 21, p. 101600, 2019.
[62] A. D. Barber *et al.*, "Connectivity supporting attention in children with attention deficit hyperactivity disorder," vol. 7, pp. 68-81, 2015.
[63] A. F. J. T. J. o. p. Arnsten, "The emerging neurobiology of attention deficit hyperactivity disorder: the key role of the prefrontal association cortex," vol. 154, no. 5, p. I, 2009.
[64] L. Weyandt, A. Swentosky, and B. G. J. D. n. Gudmundsdottir, "Neuroimaging and ADHD: fMRI, PET, DTI findings, and methodological limitations," vol. 38, no. 4, pp. 211-225, 2013.
[65] A. dos Santos Siqueira, B. Junior, C. Eduardo, W. E. Comfort, L. A. Rohde, and J. R. Sato, "Abnormal functional resting-state networks in ADHD: graph theory and pattern recognition analysis of fMRI data," *BioMed Research International*, vol. 2014, 2014.
[66] A. Ghanizadeh, M. Bahrani, R. Miri, and A. Sahraian, "Smell identification function in children with attention deficit hyperactivity disorder," *Psychiatry investigation*, vol. 9, no. 2, p. 150, 2012.
[67] A. M. Herman, H. Critchley, and T. J. S. r. Duka, "Decreased olfactory discrimination is associated with impulsivity in healthy volunteers," vol. 8, no. 1, p. 15584, 2018.
[68] J. E. Bailes, A. L. Petraglia, B. I. Omalu, E. Nauman, and T. Talavage, "Role of subconcussion in repetitive mild traumatic brain injury," *J Neurosurg*, vol. 119, no. 5, pp. 1235-45, Nov 2013.
[69] B. Johnson, T. Neuberger, M. Gay, M. Hallett, and S. Slobounov, "Effects of subconcussive head trauma on the default mode network of the brain," *J Neurotrauma*, vol. 31, no. 23, pp. 1907-13, Dec 1 2014.
[70] B. R. Cobb *et al.*, "Head impact exposure in youth football: elementary school ages 9-12 years and the effect of practice structure," *Ann Biomed Eng*, vol. 41, no. 12, pp. 2463-73, Dec 2013.
[71] R. Saxena *et al.*, "Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci," *Am J Hum Genet*, vol. 90, no. 3, pp. 410-25, Mar 9 2012.
[72] S. M. Slobounov *et al.*, "The effect of repetitive subconcussive collisions on brain integrity in collegiate football players over a single football season: A multi-modal neuroimaging study," *Neuroimage Clin*, vol. 14, pp. 708-718, 2017.
[73] D. C. Zhu *et al.*, "A potential biomarker in sports-related concussion: brain functional connectivity alteration of the default-mode network measured with longitudinal resting-state fMRI over thirty days," *J Neurotrauma*, vol. 32, no. 5, pp. 327-41, Mar 1 2015.
[74] B. Johnson *et al.*, "Alteration of brain default network in subacute phase of injury in concussed individuals: resting-state fMRI study," *Neuroimage*, vol. 59, no. 1, pp. 511-8, Jan 2 2012.

[75] K. Abbas *et al.*, "Effects of repetitive sub-concussive brain injury on the functional connectivity of Default Mode Network in high school football athletes," *Dev Neuropsychol*, vol. 40, no. 1, pp. 51-6, Jan 2015.
[76] Y. Zhou *et al.*, "Default-mode network disruption in mild traumatic brain injury," *Radiology*, vol. 265, no. 3, pp. 882-92, Dec 2012.
[77] T. J. O'Neill, E. M. Davenport, G. Murugesan, A. Montillo, and J. A. Maldjian, "Applications of Resting State Functional MR Imaging to Traumatic Brain Injury," *Neuroimaging Clin N Am*, vol. 27, no. 4, pp. 685-696, Nov 2017.
[78] C. Gallant and D. Good, "Investigating the relationship between subconcussion and psychiatric symptoms," *Archives of Physical Medicine and Rehabilitation*, vol. 97, no. 10, pp. e65-e66, 2016.
[79] G. Murugesan *et al.*, "Changes in resting state MRI networks from a single season of football distinguishes controls, low, and high head impact exposure," *Proc IEEE Int Symp Biomed Imaging*, vol. 2017, pp. 464-467, Apr 2017.
[80] B. Saghafi *et al.*, "Quantifying the Association between White Matter Integrity Changes and Subconcussive Head Impact Exposure from a Single Season of Youth and High School Football using 3D Convolutional Neural Networks," *Proc SPIE Int Soc Opt Eng*, vol. 10575, Feb 2018.
[81] F. Haran, J. D. Handy, R. J. Servatius, C. K. Rhea, and J. W. J. A. N. A. Tsao, "Acute neurocognitive deficits in active duty service members following subconcussive blast exposure," pp. 1-13, 2019.
[82] A. P. Lavender *et al.*, "Repeated Long-Term Sub-concussion Impacts Induce Motor Dysfunction in Rats: A Potential Rodent Model," vol. 11, p. 491, 2020.
[83] B. B. Reynolds, A. N. Stanton, S. Soldozy, H. P. Goodkin, M. Wintermark, and T. J. Druzgal, "Investigating the effects of subconcussion on functional connectivity using mass-univariate and multivariate approaches," *Brain imaging and behavior*, vol. 12, no. 5, pp. 1332-1345, 2018.
[84] G. K. Murugesan *et al.*, "BrainNET: Inference of brain network topology using Machine Learning," p. 776641, 2019.
[85] E. M. Davenport *et al.*, "Abnormal white matter integrity related to head impact exposure in a season of high school varsity football," *Journal of neurotrauma*, vol. 31, no. 19, pp. 1617-1624, 2014.
[86] J. J. Crisco *et al.*, "Frequency and location of head impact exposures in individual collegiate football players," *Journal of athletic training*, vol. 45, no. 6, pp. 549-559, 2010.
[87] J. E. Urban *et al.*, "Head impact exposure in youth football: high school ages 14 to 18 years and cumulative impact analysis," *Ann Biomed Eng*, vol. 41, no. 12, pp. 2474-87, Dec 2013.
[88] S. Rowson and S. M. Duma, "Development of the STAR evaluation system for football helmets: integrating player head impact exposure and risk of concussion," *Annals of biomedical engineering*, vol. 39, no. 8, pp. 2130-2140, 2011.
[89] S. Rowson and S. M. Duma, "Brain injury prediction: assessing the combined probability of concussion using linear and rotational head acceleration," *Annals of biomedical engineering*, vol. 41, no. 5, pp. 873-882, 2013.
[90] S. Rowson *et al.*, "Rotational head kinematics in football impacts: an injury risk function for concussion," *Annals of biomedical engineering*, vol. 40, no. 1, pp. 1-13, 2012.
[91] K. J. Friston, "Statistical parametric mapping," 1994.
[92] V. D. Calhoun, T. Adali, G. D. Pearlson, and J. Pekar, "A method for making group inferences from functional MRI data using independent component analysis," *Human brain mapping*, vol. 14, no. 3, pp. 140-151, 2001.
[93] V. Calhoun and T. Adali, "Group ICA of fMRI toolbox (GIFT)," *Online at http://icatb. sourceforge. net*, 2004.
[94] O. Dipasquale, L. Griffanti, M. Clerici, R. Nemni, G. Baselli, and F. Baglio, "High-dimensional ICA analysis detects within-network functional connectivity damage of default-mode and sensory-motor networks in Alzheimer's disease," *Frontiers in human neuroscience*, vol. 9, p. 43, 2015.
[95] M. J. McKeown, L. K. Hansen, and T. J. Sejnowsk, "Independent component analysis of functional MRI: what is signal and what is noise?," *Current opinion in neurobiology*, vol. 13, no. 5, pp. 620-629, 2003.
[96] X.-N. Zuo *et al.*, "Network centrality in the human functional connectome," *Cerebral cortex*, vol. 22, no. 8, pp. 1862-1875, 2012.
[97] E. Bullmore and O. Sporns, "The economy of brain network organization," *Nature Reviews Neuroscience*, vol. 13, no. 5, pp. 336-349, 2012.
[98] J. D. Medaglia, "Graph theoretic analysis of resting state functional MR imaging," *Neuroimaging Clinics*, vol. 27, no. 4, pp. 593-607, 2017.
[99] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825-2830, 2011.
[100] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, 1995, vol. 14, no. 2, pp. 1137-1145: Montreal, Canada.
[101] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, p. 91, 2006.
[102] H. Zetterberg *et al.*, "Head trauma in sports–clinical characteristics, epidemiology and biomarkers," vol. 285, no. 6, pp. 624-634, 2019.
[103] W. H. Thompson, E. P. Thelin, A. Lilja, B.-M. Bellander, and P. Fransson, "Functional resting-state fMRI connectivity correlates with serum levels of the S100B protein in the acute phase of traumatic brain injury," *NeuroImage: Clinical*, vol. 12, pp. 1004-1012, 2016.
[104] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.
[105] J. E. Villanueva-Meyer, M. C. Mabray, and S. J. N. Cha, "Current clinical brain tumor imaging," vol. 81, no. 3, pp. 397-415, 2017.
[106] N. Hoggard and G. H. J. T. B. j. o. r. Roditi, "T 1 hyperintensity on brain imaging subsequent to gadolinium-based contrast agent administration: what do we know about intracranial gadolinium deposition?," vol. 90, no. 1069, p. 20160590, 2017.
[107] J. Wang *et al.*, "Manganese-Based Contrast Agents for Magnetic Resonance Imaging of Liver Tumors: Structure–Activity Relationships and Lead Candidate Evaluation," vol. 61, no. 19, pp. 8811-8824, 2018.
[108] S. Viswanathan, Z. Kovacs, K. N. Green, S. J. Ratnakar, and A. D. J. C. r. Sherry, "Alternatives to gadolinium-based metal chelates for magnetic resonance imaging," vol. 110, no. 5, pp. 2960-3018, 2010.
[109] E. Gong, J. M. Pauly, M. Wintermark, and G. J. J. o. M. R. I. Zaharchuk, "Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI," vol. 48, no. 2, pp. 330-340, 2018.
[110] J. Kleesiek *et al.*, "Can virtual contrast enhancement in brain MRI replace gadolinium?: a feasibility study," vol. 54, no. 10, pp. 653-660, 2019.
[111] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," vol. 34, no. 10, pp. 1993-2024, 2014.

[112] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," vol. 4, p. 170117, 2017.
[113] T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. J. H. b. m. Pfefferbaum, "The SRI24 multichannel atlas of normal adult human brain structure," vol. 31, no. 5, pp. 798-819, 2010.
[114] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018.
[115] N. J. Tustison *et al.*, "Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements," vol. 99, pp. 166-179, 2014.
[116] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
[117] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.
[118] N. Ichimura, "Spatial Frequency Loss for Learning Convolutional Autoencoders," *arXiv preprint arXiv:1806.02336*, 2018.
[119] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, 2016, pp. 694-711: Springer.
[120] G. K. Murugesan *et al.*, "Multidimensional and Multiresolution Ensemble Networks for Brain Tumor Segmentation," p. 760124, 2019.
[121] G. K. Murugesan *et al.*, "Multidimensional and Multiresolution Ensemble Networks for Brain Tumor Segmentation," in *International MICCAI Brainlesion Workshop*, 2019, pp. 148-157: Springer.
[122] Y. Zhuge *et al.*, "Brain tumor segmentation using holistically nested neural networks in MRI images," *Medical physics*, vol. 44, no. 10, pp. 5234-5243, 2017.
[123] C. G. Bangalore Yogananda *et al.*, "A Fully Automated Deep Learning Network for Brain Tumor Segmentation," *Tomography*, vol. 6, no. 2, pp. 186-193, Jun 2020.
[124] R. Saouli, M. Akil, and R. Kachouri, "Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images," *Computer methods and programs in biomedicine*, vol. 166, pp. 39-49, 2018.
[125] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241: Springer.
[126] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1240-1251, 2016.
[127] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18-31, 2017.
[128] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*, 2018, pp. 311-320: Springer.
[129] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, p. 170117, 2017.
[130] S. Bakas *et al.*, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection," *The Cancer Imaging Archive*, vol. 286, 2017.
[131] S. Bakas *et al.*, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017)," ed, 2017.
[132] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," vol. 115, no. 13, pp. E2970-E2979, 2018.
[133] S. Kim *et al.*, "The roles of 11 C-acetate PET/CT in predicting tumor differentiation and survival in patients with cerebral glioma," vol. 45, no. 6, pp. 1012-1020, 2018.
[134] T. Tsuchida, H. Takeuchi, H. Okazawa, T. Tsujikawa, Y. J. N. m. Fujibayashi, and biology, "Grading of brain glioma with 1-11C-acetate PET: comparison with 18F-FDG PET," vol. 35, no. 2, pp. 171-176, 2008.
[135] Y. Yamamoto *et al.*, "11 C-acetate PET in the evaluation of brain glioma: Comparison with 11 C-methionine and 18 F-FDG-PET," vol. 10, no. 5, p. 281, 2008.
[136] P. Yakubovskiy, "Segmentation Models," *GitHub repository*, GitHub 2019.
[137] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.
[138] C.-F. Chen, Q. Fan, N. Mallinar, T. Sercu, and R. J. a. p. a. Feris, "Big-little net: An efficient multi-scale feature representation for visual and speech recognition," 2018.
[139] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.
[140] L. Chen, Y. Wu, A. M. DSouza, A. Z. Abidin, A. Wismüller, and C. Xu, "MRI tumor segmentation with densely connected 3D CNN," in *Medical Imaging 2018: Image Processing*, 2018, vol. 10574, p. 105741F: International Society for Optics and Photonics.
[141] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. J. I. t. o. m. i. Ayed, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," vol. 38, no. 5, pp. 1116-1126, 2018.
[142] J. Islam and Y. J. a. p. a. Zhang, "An Ensemble of Deep Convolutional Neural Networks for Alzheimer's Disease Detection and Classification," 2017.
[143] J. J. Van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," vol. 77, no. 21, pp. e104-e107, 2017.
[144] G. R. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. J. J. O. S. S. O'Leary, "PyWavelets: A Python package for wavelet analysis," vol. 4, no. 36, p. 1237, 2019.
[145] X. Feng, N. Tustison, and C. Meyer, "Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features," in *International MICCAI Brainlesion Workshop*, 2018, pp. 279-288: Springer.
[146] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," vol. 12, no. Oct, pp. 2825-2830, 2011.
[147] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," 2013.
[148] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794: ACM.
[149] S. Nalawade *et al.*, "Classification of brain tumor isocitrate dehydrogenase status using MRI and deep learning," *J Med Imaging (Bellingham)*, vol. 6, no. 4, p. 046003, Oct 2019.
[150] D. W. Parsons *et al.*, "An integrated genomic analysis of human glioblastoma multiforme," *Science*, 2008.
[151] H. Yan *et al.*, "IDH1 and IDH2 mutations in gliomas," *New England Journal of Medicine*, vol. 360, no. 8, pp. 765-773, 2009.

[152] D. N. Louis *et al.*, "The 2016 World Health Organization classification of tumors of the central nervous system: a summary," *Acta neuropathologica*, vol. 131, no. 6, pp. 803-820, 2016.
[153] W. B. Pope *et al.*, "Non-invasive detection of 2-hydroxyglutarate and other metabolites in IDH1 mutant glioma patients using magnetic resonance spectroscopy," *Journal of neuro-oncology*, vol. 107, no. 1, pp. 197-205, 2012.
[154] C. Choi *et al.*, "2-hydroxyglutarate detection by magnetic resonance spectroscopy in IDH-mutated patients with gliomas," *Nature medicine*, vol. 18, no. 4, p. 624, 2012.
[155] M. I. de la Fuente *et al.*, "Integration of 2-hydroxyglutarate-proton magnetic resonance spectroscopy into clinical practice for disease monitoring in isocitrate dehydrogenase-mutant glioma," *Neuro-oncology*, vol. 18, no. 2, pp. 283-290, 2015.
[156] A. Tietze *et al.*, "Noninvasive assessment of isocitrate dehydrogenase mutation status in cerebral gliomas by magnetic resonance spectroscopy in a clinical setting," *Journal of neurosurgery*, vol. 128, no. 2, pp. 391-398, 2017.
[157] C. Choi *et al.*, "Prospective longitudinal analysis of 2-hydroxyglutarate magnetic resonance spectroscopy identifies broad clinical utility for the management of patients with IDH-mutant glioma," *Journal of Clinical Oncology*, vol. 34, no. 33, p. 4030, 2016.
[158] C. Choi *et al.*, "A comparative study of short-and long-TE 1H MRS at 3 T for in vivo detection of 2-hydroxyglutarate in brain tumors," *NMR in Biomedicine*, vol. 26, no. 10, pp. 1242-1250, 2013.
[159] S. K. Ganji *et al.*, "In vivo detection of 2-hydroxyglutarate in brain tumors by optimized point-resolved spectroscopy (PRESS) at 7T," *Magnetic resonance in medicine*, vol. 77, no. 3, pp. 936-944, 2017.
[160] R. L. Delfanti *et al.*, "Imaging correlates for the 2016 update on WHO classification of grade II/III gliomas: implications for IDH, 1p/19q and ATRX status," *Journal of neuro-oncology*, vol. 135, no. 3, pp. 601-609, 2017.
[161] K. Chang *et al.*, "Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging," *Clin Cancer Res*, vol. 24, no. 5, pp. 1073-1081, Mar 1 2018.
[162] X. Zhang *et al.*, "Radiomics Strategy for Molecular Subtype Stratification of Lower-Grade Glioma: Detecting IDH and TP53 Mutations Based on Multimodal MRI," *J Magn Reson Imaging*, Feb 2 2018.
[163] P. Chang *et al.*, "Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas," *AJNR Am J Neuroradiol*, vol. 39, no. 7, pp. 1201-1207, Jul 2018.
[164] Z. Akkus *et al.*, "Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence," *J Digit Imaging*, vol. 30, no. 4, pp. 469-476, Aug 2017.
[165] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045-1057, 2013.
[166] L. Scarpace *et al.*, "Radiology data from the cancer genome atlas glioblastoma multiforme [tcga-gbm] collection," *The Cancer Imaging Archive*, vol. 11, p. 4, 2016.
[167] N. Pedano, A. Flanders, and L. Scarpace, "Radiology data from The Cancer Genome Atlas Low Grade Glioma [TCGA-LGG] collection," *Cancer Imaging Arch*, 2016.
[168] M. Ceccarelli *et al.*, "Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma," (in eng), *Cell*, vol. 164, no. 3, pp. 550-63, Jan 28 2016.
[169] X. Feng, J. Yang, Z. C. Lipton, S. A. Small, and F. A. Provenzano, "Deep Learning on MRI Affirms the Prominence of the Hippocampal Formation in Alzheimer's Disease Classification," *bioRxiv*, p. 456277, 2018.
[170] D. A. Bluemke, "Editor's Note: Publication of AI Research in Radiology," *Radiology*, vol. 289, no. 3, pp. 579-580, Dec 2018.
[171] C. Szegedy, "Inception-v4, inception-resnet and the impact of residual connections on learning.," *AAAI*, vol. Vol. 4. 2017, 2017.
[172] C. Szegedy, "Going deeper with convolutions.," *IEEE Transactions on Computational Imaging*, 2015.
[173] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*, 2017, vol. 1, no. 2, p. 3.
[174] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[175] H. Robbins and S. J. T. a. o. m. s. Monro, "A stochastic approximation method," pp. 400-407, 1951.
[176] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 116: ACM.
[177] V. Wegmayr, S. Aitharaju, and J. Buhmann, "Classification of brain MRI with big data and deep 3D convolutional neural networks," in *Medical Imaging 2018: Computer-Aided Diagnosis*, 2018, vol. 10575, p. 105751S: International Society for Optics and Photonics.