Enhancing Medical Imaging Workflows with Deep Learning

by

Ken Chang

M.S.E. Bioengineering
University of Pennsylvania, 2013

SUBMITTED TO THE DEPARTMENT OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MEDICAL ENGINEERING AND MEDICAL PHYSICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Signature of Author: ____

Department of Health Sciences and Technology
May 28, 2020

Certified by: ____

Jayashree Kalpathy-Cramer, PhD
Associate Professor of Radiology
Thesis Supervisor

Certified by: ____

Bruce R. Rosen, MD, PhD
Professor of Health Sciences and Technology
Thesis Supervisor

Accepted by: ____

Emery N. Brown, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology
Professor of Computational Neuroscience and Health Sciences and Technology

Enhancing Medical Imaging Workflows with Deep Learning

by

Ken Chang

Submitted to the Department of Health Sciences and Technology
on May 28, 2020 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Medical Engineering and Medical Physics

ABSTRACT

The last few years mark a significant leap in the capability of algorithms with the advent of deep learning. While conventional machine learning has existed for decades, their utility has been rather limited, requiring considerable engineering and domain expertise to design pertinent data features that can be extracted from raw data. In contrast, deep learning methods have yielded state-of-the-art results in a wide range of computer vision tasks without the need for hand-crafted imaging features. At the same time, we are collecting ever-increasing quantities of medical imaging. Together, deep learning models and big data yield a powerful combination. Integrated in the data workflow, the clinic, or at the bedside, these models have the potential to aid with clinical decision-making, improving efficiency, accuracy, and reliability of patient care. However, at present, there is a critical gap between the researchers who develop deep learning algorithms and the clinicians who could utilize the technology to improve patient care. In this thesis, I focus on several challenges that prevent clinical translation of algorithms. First, vast quantities of data needed to train effective models are often dispersed across institutions and cannot be shared due to ethical, infrastructure, and patient privacy concerns. As such, we developed distributed methods of training robust deep learning models that do not require sharing patient data in multi-institutional collaborative settings. Second, it is not clearly understood how decisions in algorithm design can affect model performance. To this end, I showcase how various training, data, and model parameters can impact algorithm prediction and performance. Lastly, while many algorithms are designed to perform a single task, there are few pipelines that have multi-faceted functionality needed in patient care. I demonstrate an integrated and deployable clinical decision support pipeline for glioma and ischemic stroke that is extensible to other diseases.

Thesis Supervisor: Jayashree Kalpathy-Cramer, PhD
Title: Associate Professor of Radiology, HMS

Thesis Supervisor: Bruce R. Rosen, MD, PhD
Title: Professor of Health Sciences and Technology, HMS

Acknowledgements

For this body of work, I am indebted to many. First off, I would like to thank my thesis advisors, Jayashree Kalpathy-Cramer and Bruce Rosen. Their knowledge, expertise, and experience have been a light on this long journey, without which I would have been lost at many points. Their dedication to their students has inspired me to one day do the same for other students. Their mentorship goes far beyond research, as they have taught me of what it means to truly care for your students and how to develop professionally and personally. Their teachings have taught me to overcome daily hurdles without losing sight of the big picture. Importantly, I have learned from them how to communicate and collaborate as well as how to be fearless and persistent in my pursuits ("*Don't ask, don't get*" as Bruce likes to say), and not to be discouraged by inevitable setbacks. Further, I learned from Jayashree that spice can be added to everything (research, writing, food, *dessert*, and otherwise) and there is no such thing as too spicy (even when your mouth is on fire).

I also want to thank my committee members, Elfar Adalsteinsson and Bruce Fischl, for providing positive and diverse perspectives as well as shaping my research to new directions. I am truly grateful and incredibly lucky to have their seasoned expertise. I also want to thank the many members of the QTIM lab over the years, who have fostered a productive, creative, collaborative, and fun environment. I want to give a special shout-out to Elizabeth Gerstner, who has provided many meaningful interactions and teachings, both inside and outside of the clinic, as well as critical contributions to my research and writing (as well as many laughs within the lab). The other members to thank include James Brown, Andrew Beers, Malika Shahrawat, Samarth Nandekar, Yi-Fen Yen, Ina Ly, Katharina Hoebel, Jay Patel, Praveer Singh, Matthew Li, Jonathan Cardona, Sunakshi Paul, Benjamin Bearce, Kevin Lou, Hyunji Kim, Dania Daye, Bryan Chen, Sean Ko, Nishanth TA, Nathan Gaw, Xiaoyue Ma, Alton Sartor, Albert Kim, Mishka Gidwani, Ikbeom Jang, Mehak Aggarwal, Sharut Gupta, and Witwisit Kantaprom. The QTIM community is indeed a special one that I will relish for the rest of my life. The same applies to the NTP, Martinos, and HST community, who have fostered a fun and intellectual environment.

On my professional journey, I have also had many other medical school, graduate school, and training program advisors who I want to thank: Randy Gollub (and her always positive and energetic persona), Ralph Weissleder, Allan Goldstein, Rick Mitchell, Matthew Frosch, Julie Greenberg, Loren Walensky, and Mohini Lutchman. On the administrative side, I want to thank Laurie Ward, Joe Stein, Patty Cunningham, and Amy Cohen for helping me navigate the many seemingly endless academic requirements needed for the dual degree program.

During the course of my graduate research, I have also had the privilege of collaborating with many researchers from many institutions. These collaborations have granted me opportunities beyond my wildest dreams and for that I am thankful. These collaborators include those from Stanford University (Niranjan Balachandar, Darvin Yi, Daniel Rubin), Brigham and Women's Hospital (Raymond Huang, Patrick Wen, Omar Arnaout, Joeky Senders, Vasileios Kavouridis, Alessandro Boaro, Wenya Linda Bi), University of Pennsylvania (Chang Su, Harrison Bai), Massachusetts General Hospital (Otto Rapalino, Shirley Chou, Sadia Choudhery, Patricia Musolino, Maarten Poirot, Rajiv Gupta, Gil Gonzalez), Martinos Center (Hakan Ay, Angel Torrado Carvajal, Marco Loggia, Eva-Maria Ratai, Daniel Kim, Michael Wenke), Center for Clinical Data Science (Ram Naidu, Katherine Andriole, Nir Neumark, Christopher Bridge, Bernardo Bizzo, Romane Gauriau), American College of Radiology (Laura Coombs, Mike

Tilkin, Etta Pisano, Laura Coombs, Keith Dreyer, Bibb Allen, Sheela Agarwal), NVIDIA (Holger Roth, Wenqi Li, Alvin Ihsani), MIT Media Lab (Praneeth Vepakomma, Ramesh Raskar), and A*STAR (Pavitra Krishnaswamy). I want to especially thank Raymond Huang who has been a constant source of advice and perspective during graduate school. I also wanted to acknowledge NIH for their research support and the Center for Clinical Data Science for providing computational resources, without which this research would not be possible.

I also want to thank my undergraduate research advisor, Tobias Baumgart, and my late research advisor at the NIH, Alexander Gorbach, for inspiring me to pursue graduate studies and an academic career. Although Dr. Gorbach is no longer with us, his memory and teachings reside with me, and for that, I will forever be grateful. I also want to thank Benjamin Capraro, who as a graduate student, patiently mentored me day-to-day when I was a clueless undergrad. Finally, I want to give thanks to my friends, my parents, my brother, and my partner Alice Choi for their constant encouragement and support. I couldn’t imagine embarking on this journey without them.

Permissions

I have affirmation from Oxford University Press that I can reproduce the following manuscripts in this dissertation:

Chang, K.*, Balachandar, N.*, Lam, C., Yi, D., Brown, J., Beers, A., Rosen, B., Rubin, D.L., and Kalpathy-Cramer, J., 2018. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association*, 25(8), pp.945-954. DOI: 10.1093/jamia/ocy017. PMID: 29617797

Chang, K.*, Beers, A.L.*, Bai, H.X.*, Brown, J.M., Ly, K.I., Li, X., Senders, J.T., Kavouridis, V.K., Boaro, A., Su, C., Bi, W.L., Rapalino, O., Liao, W., Shen, Q., Zhou, H., Xiao, B., Wang, Y., Zhang, P.J., Pinho, M.C., Wen, P.Y., Batchelor, T.T., Boxerman, J.L., Arnaout, O., Rosen, B.R., Gerstner, E.R., Yang, L., Huang, R.Y., and Kalpathy-Cramer, J., 2019. Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bi-dimensional measurement. *Neuro-Oncology*. DOI:10.1093/neuonc/noz106. PMID: 31190077

The following manuscript was reproduced under AACR usage guidelines:

Chang, K.*, Bai, H.X.*, Zhou, H., Su, C., Bi, W.L., Agbodza, E., Zhang, B., Capellini, A., Liao, W., Shen, Q., Li, X., Xiao, B., Cryan, J., Ramkissoon, S., Ramkissoon, L., Ligon, K., Wen, P.Y., Bindra, R., Woo, J., Arnaout, O., Gerstner, E.R., Zhang, P.J., Rosen, B.R., Yang, L., Huang, R.Y., Kalpathy-Cramer, J., 2018. Residual convolutional neural network for the determination of IDH status in low-and high-grade gliomas from MR imaging. *Clinical Cancer Research*, 24(5), pp.1073-1081. DOI:10.1158/1078-0432.CCR-17-2236. PMID: 29167275

The following manuscripts are currently under review:

Chang, K., Beers, A.L., Brink, L., Patel, J.B., Singh, P., Arun, N.T., Hoebel, K.V., Gaw, N., Shah, M., Pisano, E.D., Tilkin, M., Coombs, L.P., Dreyer, K.J., Allen, B., Agarwal, S., Kalpathy-Cramer, J., 2020. Multi-Institutional Assessment and Crowdsourcing Evaluation of Deep Learning for Automated Classification of Breast Density.

Contents

List of Figures

List of Tables

1 Introduction

The last few years mark a significant leap in the capability of classification, detection, and segmentation algorithms with a new class of techniques under the umbrella of deep learning. While conventional machine learning has existed for decades, their utility has been rather limited, requiring considerable engineering and domain expertise to design pertinent data features that can be extracted from raw data. In contrast, deep learning methods do not require domain-inspired hand-crafted imaging features. With the advent of powerful graphics processing units, deep learning has brought about major breakthroughs in tasks such as image classification, speech recognition, and natural language processing.[1–3] Deep learning models take raw data such as images as input and apply many layers of transformations to calculate the output of interest. The high dimensionality of these transformations allows these algorithms to learn complex patterns with a high level of abstraction.[4] The logical application of deep learning, especially those methods developed for computer vision, is to medical imaging, where clinicians have long noticed the relationship between imaging patterns and diagnoses, prognosis, and genomics.

At the same time, we are collecting ever-increasing quantities of medical imaging. Together, deep learning models and big data yield a powerful combination. Integrated in the data workflow, the clinic, or at the bedside, these models have the potential to aid with clinical decision-making, improving efficiency, accuracy, and reliability of patient care. However, at present, there is a critical gap between the scientists who develop deep learning algorithms and the clinicians who will utilize the technology to improve patient care. Our long-term objective is to develop artificial intelligence technologies for medical imaging that can improve the efficiency, cost, and quality of patient care.

In this thesis, I focus on several challenges that prevent clinical translation of algorithms. First, vast quantities of data needed to train effective models are often dispersed across institutions and cannot be shared due to ethical, infrastructure, patient privacy concerns. As such, we developed distributed methods of training deep learning models that do not require sharing patient data in multi-institutional collaborative settings. Second, it is not clearly understood how decisions in algorithm design can affect model performance. In our work, we showcase how various training, data, and model parameters can impact algorithm prediction and performance. Lastly, while many algorithms are designed to perform a single task, there are few pipelines that have the multi-faceted functionality needed in patient care. We demonstrate an integrated and deployable clinical decision support platform for glioma and ischemic stroke that is extensible to other diseases.

1.1 Distributed deep learning networks among institutions as an alternative to sharing patient data

1.1.1 Motivation

Deep learning has become a promising approach for automated support for clinical diagnosis. Deep learning is most effective when trained on large, diverse datasets. However, when medical data samples are limited, collaboration among multiple institutions is necessary to achieve high algorithm performance. Sharing patient data often has limitations due to technical, legal, or ethical concerns. In chapter 3, we propose methods of distributing deep learning models as an attractive alternative to sharing patient data.

1.1.2 Contributions

We simulate the distribution of deep learning models across four institutions using various training heuristics and compare the results with a deep learning model trained on centrally hosted patient data.[5,6] The training heuristics investigated include ensembling single institution models, single weight transfer, and cyclical weight transfer. We evaluated these approaches for image classification in three independent image collections (retinal fundus photos, mammography, and ImageNet). Among the results, we find that:

- High model performance can be achieved without centrally hosted data. Distributing deep learning can effectively utilize data from many institutions as long as the institutions are willing to share the model parameters. Specifically, we show that cyclical weight transfer resulted in a performance that was comparable to that of centrally hosted patient data.
- There is an improvement in the performance of cyclical weight transfer heuristic with high frequency of weight transfer, which represents a tradeoff between communication costs and performance.
- Performance of cyclical weight transfer is compromised in the presence of data quality and data size variability at a single institution. We also propose methods to optimize distributed learning in the presence of data size and label imbalance heterogeneity across multiple institutions.[7]
- Distributed deep learning models can achieve high performance at large scale (that is, when there are many institutions, each with small amounts of data).

1.2 Model design and the impact on performance

1.2.1 Motivation

Despite the significant progress in research, AI is currently underutilized in current clinical workflows. The success of AI model development depends critically on the synergy between the availability of high-quality datasets, physicians who can drive clinical direction, and data scientists who can design effective algorithms. Even after the model is developed, aspects such as clinician misunderstanding of model limitations, limited model generalizability due to lack of continual refinement and collaborative training, and lack of rigorous validation can preclude integration into the clinical workflow. In chapter 4, we explore these issues.

1.2.2 Contributions

Accurate and consistent evaluation of mammographic breast density is an unmet clinical need. We develop deep learning algorithms to assess Breast Imaging Reporting and Data System breast density, investigating the effect of data, model, and training parameters on overall model performance. For our study, we utilized a large multi-institution patient cohort of 108,230 digital screening mammograms from the Digital Mammographic Imaging Screening Trial. Our best performing algorithm achieved good agreement with radiologists, with a 4-class κ of .667. Among the technical results, we find that:

- ImageNet pretraining of models improved performance compared to randomly initialized models.
- Model performance increased with increasing training set size. However, this increase reaches a plateau with larger quantities of training data.

- Model architecture, ensembling, and augmentation had smaller effects on model performance.
- Randomly sampling of images at each training iteration can significantly bias model predictions away from low-represented classes when compared to using sampling each density class with equal probability. The net result is an increase in sensitivity and a decrease in specificity for predicting dense breasts for equal class compared to random sampling.
- Performance of the model degrades when we evaluate on digital mammography data formats that differ from the one that we trained on, emphasizing the importance of multi-institutional training sets.
- In exploring the learned features of the algorithm, we discover that the algorithm learns a distribution for each data format as opposed to a unified distribution across all data formats.
- When the model was fine-tuned on data from a new institution, the model had lower performance on the original dataset it was trained on, a phenomenon known as catastrophic forgetting.
- We provide a crowdsourcing evaluation from the attendees of the American College of Radiology 2019 Annual Meeting, showing that crowdsourced annotations, including those from attendees who routinely read mammograms, have higher agreement with our algorithm than with interpreting radiologists. Because crowdsourced annotations are minimally time-consuming for individual participants, it provides an effective way to evaluate algorithms.

- Because our results have implications for any deep learning medical imaging study, these study design options have been integrated in the ACR AI-LAB, a platform for democratizing Artificial Intelligence that brings together clinicians and researchers to build and evaluate deep learning models.

1.3 Developing a integrated deep learning pipeline for glioma

1.3.1 Motivation

Gliomas are primary central nervous system tumors with variable natural histories and prognoses depending on their histologic and molecular characteristics. Current clinical evaluation and treatment approaches contain manual inefficiencies that can be enhanced by deep learning, specifically delineation of tumor boundaries, treatment response assessment, and non-invasive prediction of molecular markers. This is the focus of chapter 5.

1.3.2 Contributions

We create a multifaceted deep learning pipeline for glioma using a multi-institutional patient cohort from Brigham and Women’s Hospital, Massachusetts General Hospital, Hospital of University of Pennsylvania, and TCIA.

- Taking advantage of biological context, we create an integrated tool for brain extraction, fluid attenuated inversion recovery (FLAIR) hyperintensity segmentation, and contrast-enhancing tumor segmentation.[8] We show that our method for brain extraction outperforms previous methods, which do not perform well in the presence of varied image acquisition settings and disease pathology. To facilitate utilization of the brain extraction and segmentation algorithm by the larger research and clinical community, we have made the trained model publicly available as part of our open-source neuroimaging package, DeepNeuro.[9,10]

- We automatically quantitate tumor volume and bi-directional measurements as per Response Assessment in Neuro-Oncology (RANO) criteria.[11] These measures serve as automatic measures of tumor burden. We show that automatic tumor volume and AutoRANO are highly repeatable and show good agreement with clinical experts.
- Notably, automatic tumor volume and AutoRANO capture longitudinal changes in tumor burden, which serve as the basis of treatment response assessment. This tool may be helpful in clinical trials and clinical practice to decrease the time expended by clinicians for manual annotation as well as decreasing interobserver variability.
- We also developed a tool to noninvasively predict Isocitrate Dehydrogenase (*IDH)* status from MR imaging.[12] *IDH* status is of clinical importance as patients with *IDH*-mutated tumors have longer overall survival than their *IDH*-wild-type counterparts. In addition, knowledge of IDH status may guide surgical planning. By using a large, multi-institutional patient data set with a diversity of acquisition parameters, we show the potential of the approach in clinical practice. Furthermore, this algorithm offers broad applicability by utilizing conventional MR imaging sequences. Our model has potential to complement surgical biopsy and histopathologic analysis by offering molecular marker information at the time of imaging.

1.4 Developing a computational pipeline for ischemic stroke

1.4.1 Motivation

Cerebrovascular disease is the third leading cause of death around the world after heart disease and cancer.[13,14] The most common clinical manifestation of cerebrovascular disease is an acute stroke, 87% of which are of an ischemic nature.[14] In chapter 6, we investigate the inefficiencies in the current clinical evaluation approach that can be improved with computational tools, specifically delineation of stroke boundaries, quantification of stroke volumes, and prediction of symptoms that the patient can develop during the hospital course.

1.4.2 Contributions

We create a multifaceted pipeline for imaging of ischemic stroke, using a large patient cohort from the Massachusetts General Hospital.

- We developed a deep learning tool for ischemic stroke volumetric segmentation utilizing only DWI imaging. To improve segmentation performance, we modified the U-Net neural network architecture as well as ensembled the output of several trained models. We also clinical experts qualitatively assess the automatically delineated stroke boundaries, which found that clinical experts rated automatic segmentations to have equivalent or higher quality compared to manual segmentations.
- There was high agreement between manually and automatically derived volumes. Automatic volumes were capable of differentiating 90-day disability (modified Rankin Scale >2, $p < .001$) as well as 90-day survival ($p < .001$).
- We also aimed to identify the neuroanatomic correlates of a broad range of cardiac and systemic alterations occurring after ischemic stroke. Using a mapping technique that is

free from the bias of a-priori hypothesis as to any specific location, we show that both cardiac and systemic abnormalities occurring after stroke map to specific infarct locations on diffusion-weighted MR. We show that these maps are predictive of the abnormalities as well as patient outcomes.

- To facilitate further utilization, development, and validation of the segmentation algorithm by the larger research and clinical community, we have made the code as well as trained model publicly available as part of our open-source neuroimaging package, DeepNeuro.[9,10]

2 Background

2.1 The disruption of deep learning

In an increasingly digital world, it is difficult to imagine an aspect of life where computers do not play a role. The last few years mark a significant leap in the capability of classification, detection, and segmentation algorithms with a new class of techniques under the umbrella of deep learning. While conventional machine learning has existed for decades, their utility has been rather limited, requiring considerable engineering and domain expertise to design pertinent data features that can be extracted from raw data.[15] In contrast, deep learning methods have yielded state-of-the-art results in a wide range of computer vision, speech recognition, and natural language processing tasks without the need for hand-crafted features.[15–17]

Advances in technology have been spurred by large scale competitions such as ImageNet, which have brought about innovations in training and structure of these artificial intelligence (AI) algorithms.[16] At the core of deep learning are convolutional neural networks (CNNs); a machine learning technique that can be trained on raw data to predict the outputs of interest via a supervised approach. This is achieved through many layers of non-linear transforms that are capable of learning complex patterns with a high level of abstraction.[4] The abstractions can represent low-level features such as edges to high-level motifs such as the ear of a cat. With the advent of more powerful graphics processing units (GPUs) that allow for training of large-scale neural network architectures, deep learning has become the method of choice for automating tasks from images, speech, and text.

2.2 The promise for healthcare applications

The promise of deep learning has spread like a wave across numerous domains with many of them converging within medicine.[18]

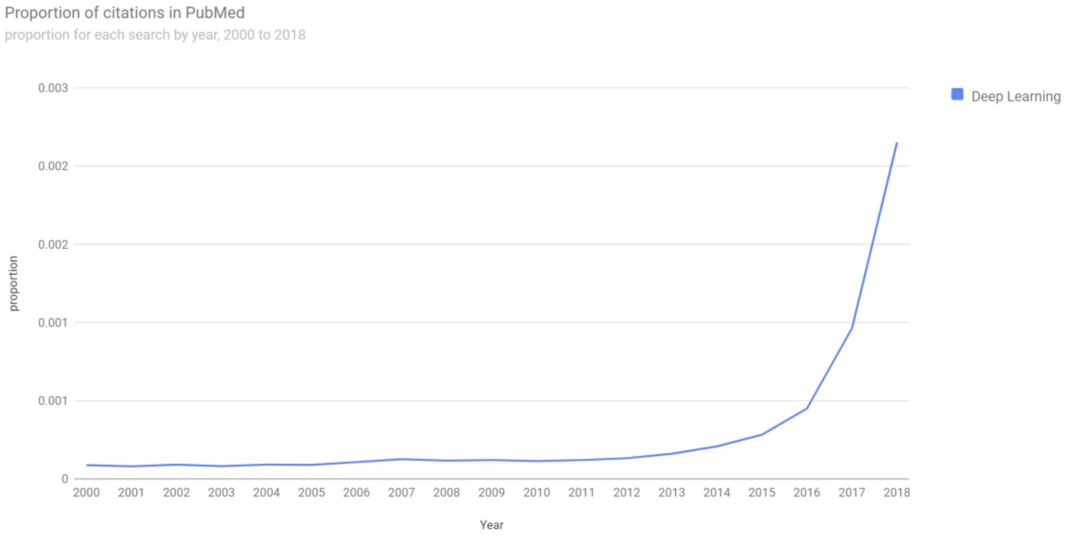


Figure 2-1. Trend of publications on PubMed that contain keyword “Deep Learning”.[19]

The timing could not be better as we are simultaneously in the era of big data. In the medical context, we are collecting ever-increasing quantities of data, in many forms including medical imaging, physiological measures, genomic sequencing, sensor data, electronic health records.[18,20–25] Conventional approaches to medical data often require domain-expertise from clinicians and medical researchers, which is then followed by formulation of biological, pathological, physiological, and/or anatomical features.[26–33] This approach is limited in many aspects because it is 1) time-consuming and challenging (as it is not always easy to engineer useful data features), 2) not scalable as there are numerous features to model, especially when trying to integrate of multi-faceted data (such as electronic health record, sequencing, imaging, sensor data), and 3) limited if the domain-knowledge is not yet known (such as determining sex or anemia from retinal fundus photographs[34,35]).

To scale into the future of medical data analysis, a less hand-crafted approach becomes a necessity, that is deep learning. However, in isolation, deep learning and big data cannot do much. Together, they yield a powerful combination. Recent studies have shown the potential of deep learning in medical fields such as dermatology, ophthalmology, and radiology for key clinical assessments, such as diagnosis, prognosis, longitudinal change detection, response to treatment, and future disease progression.[36–47] Even before the data reaches clinical personnel, deep learning has shown potential for image reconstruction, data quality assessment, and motion detection.[48–50] Integrated in the data workflow, the clinic, or at the bedside, these models have the potential to aid with clinical decision-making, improving efficiency, accuracy, and reliability of patient care.

2.3 The clinician as an information specialist

The modern clinician wears two hats: one as a healthcare provider who takes a holistic, humanistic approach to care. The other as an information specialist, integrating clinical context and results from diagnostic imaging as well as testing to make informed clinical decisions. Due to the complexity of the latter task, there are many pitfalls to have a human perform this task: namely it is time-consuming, expensive, subject to variability, and prone to human error. Elaborating further, it can be incredibly time-consuming to interpret diagnostic tests, adding expenses to already ballooning healthcare costs. A complex CT scan can take up to an hour to read.[51] This is further compounded by the increased utilization of medical imaging as well and improved technology (such as higher resolution images), further increasing workload.[20] The time spent interpreting diagnostic tests is expensive, both monetarily and for certain medical specialties at the opportunity cost of spending more time with the patient. Additionally, interpretation by humans can be highly variable. Despite attempts at objective guidelines for assessing disease, each physician's interpretation is different, which may have profound impacts on downstream decision-making.[52–54] Lastly, a hallmark of being human is to err. As medical decision making is in many regards, a manual task, it is also prone to human error. The question that much of the research in healthcare AI tries to address is whether deep learning algorithms can alleviate these pitfalls. While some have touted that machines will replace humans, that in my opinion, is much further down the line and is not even the goal of such technologies. Rather, these algorithms can bring much more value in the immediate term as a support system, augmenting human intelligence and capabilities. It has been yet to be shown that the sum of humans and machines is greater than either alone.

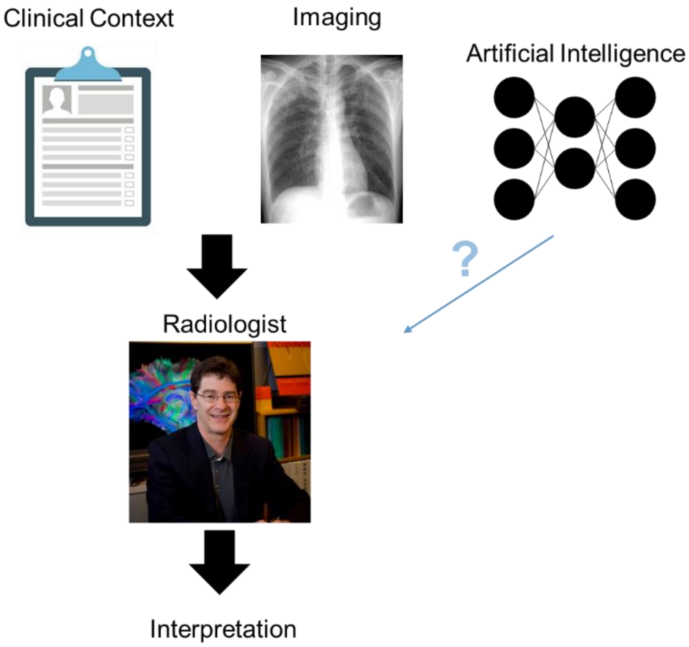


Figure 2-2 The modern clinician (for example radiologist shown here) is an information specialist. Whether AI algorithms can augment the clinician remains an open question.

2.4 The pitfalls of radiomics and other handcrafted features

Clinicians have long noticed a relationship between imaging features and genomics, prognosis, and treatment response. The term “radiomics” describes the technique of calculating various imaging features that describe shape, intensity, and texture, among many other characteristics.[55] Combined with traditional machine learning techniques (such as random forest), radiomics has been successful in segmentation as well as predicting various measures such as genomics, likelihood of metastasis, drug response, and prognosis.[56–60] However, radiomics has the limitation of relying on the computation of a selection of manually formulated or “handcrafted” features, which may not capture the full range of the information contained within imaging.

In addition, there are many challenges to the reproducibility of radiomics. While there are explicit mathematical definitions of radiomic features, differences in implementations across software packages can result in significantly different feature values.[61,62] In addition, how the radiomics package treats the boundary of the region of interest (whether through masking, omitting boundary pixels, or dilation), performs normalization of texture features, handles 3D images (as either a single 3D volume or a stack of 2D images), and quantizes pixel values can have downstream effects.[63] There is also variability among users in how they perform the segmentation (manual vs automatic) and how they pre-process the images (choice of resampling and interpolation).[63] The scanner model, acquisition parameters, and reconstruction kernel can also have an effect.[64–66] Taken together, many studies have show that radiomic features (and their resulting predictive models) are difficult to reproduce.[61,64,65,67–69] Radiomic signatures also have a number of vulnerabilities, mostly notably the lack of sensitivity to voxel randomization.[70] While a deep learning model trained on diverse training data may not mitigate all of these challenges, it does represent a less engineered approach and the next generation of automated methods. Indeed, a recent study found that deep learning outperformed radiomics for classification of contrast-enhancing lesions on multiparametric breast MR.[71] Similarly, deep learning approaches performed significantly better than texture feature-based approaches in the CAMELYON challenge for detecting lymph node metastases in women with breast cancer on pathology slides.[72]

2.5 The challenge of developing robust deep learning models

While the promise of clinical deep learning models is high, there are numerous obstacles to training effective deep learning models. First, there is a need for enormous quantities of annotated training data, especially for diseases with subtle or diverse phenotypes. The data requirement is also increased when the individual patient data are noisy or incomplete. While the vast majority of medical diseases have no publicly available datasets, the few of those that do are either limited in quantity (varying from few hundreds to hundreds of thousands) or are incompatible for transfer learning on a different medical problems.[73,74] Comparatively, most of the state-of-the-art neural network architectures have millions of parameters and have been trained on benchmark datasets such as ImageNet which have millions of annotated images.[16] In such a scenario, where publicly available dataset is scarce for a given medical problem, algorithm developers have to rely on their own institutional datasets. However, for rare diseases or when studying the effect of different modes of treatment that may be hospital specific, it may be impossible to acquire sufficient quantities of training data at a single institution or the trained model might not be generic enough to perform well on outside institution datasets.[75] Furthermore, deep learning algorithms are prone to overfitting and brittle when evaluated on external data. As such, training data needs to be diverse, ideally from varying acquisition settings and patient populations. Unfortunately, data from a single institution are often limited in quantity and heterogeneity, rendering the data insufficient for training deep learning algorithms. In such cases, multi-institutional patient cohorts are the only avenue towards training an effective deep learning model.

One approach to multi-institutional studies is to build a large central repository, but this is hindered by concerns about data sharing, specifically patient privacy, data de-identification, regulation, intellectual property, data storage. Firstly, protecting patient privacy is of upmost importance in the increasingly digital world as the release of sensitive patient information would be harmful. Recently studies have shown that the barrier to re-identification is quite low, requiring just a few clinical variables or a single scan, emphasizing the importance of privacy preservation.[76,77] Second, it is difficult to ensure rigorous patient de-identification and there is the potential of accidental data leakage. Also, data is a valuable resource and many hospitals prefer not to publicly share data to protect their own institutional interests. Lastly, patient data is growing in size with the increasing resolution and number of imaging modalities. As such, it would be cumbersome to commission the substantial data storage required to centrally host data. These challenges have made centrally hosting data both expensive and impractical. An alternative approach is to have the data be locally hosted and have the model be trained in a distributed fashion. Comparatively, the model is much smaller than patient data so the communication overhead is drastically reduced. In a recent study, we showed that distributed/federated approaches can achieve centrally hosted performance without sharing patient data [5]. Under this paradigm, each institution will install a software application that links the different institutions together, allowing for collaboration and distribution computation. The result is a model that performs as well as if the data had been shared, while still preserving patient privacy and protecting institutional interests. In addition, the reduced requirements for storage and de-identification means reduced cost of collaboration and increased incentive for participation in multi-institutional studies.

One critical hurdle that prevents the deployment of deep learning models in the clinical work environment is their relatively poor generalizability across institutional differences, such as patient demographics, disease prevalence, scanners, and acquisition settings. A variety of recent deep learning studies that have shown poor generalizability of deep learning models when applied to data from different institutions than the one they were trained on [78,79]. The optimal method of distributing the process of training deep learning models across heterogenous institutions has not yet been studied.

2.6 Understanding design and limitations of deep learning models

Despite significant research into the applications of AI, there is currently limited use of AI in clinical care. Key to the success of these algorithms are two components: 1) clinical professionals who can drive direction, validation, and translation, and 2) data scientists who can design, train, and deploy such algorithms. Unfortunately, such synergy is only accessible within certain academic institutions. To date, physicians, who have the requisite domain expertise to make AI relevant for clinical use, have not been able to widely participate in AI development because of limited access to AI computational solutions and education resources. On the other hand, data scientists who are not working closely with radiology professionals are building algorithms that are accurate yet clinically irrelevant or not useful.

Even once the model is trained, there are many foregoing challenges. First, end-users for these models, such as radiologists and pathologists, may not fully understand how these models were trained. Critically, this may lead to a general misunderstanding for why and when a model will fail. Furthermore, models need to be rigorously validated to ensure that they are not biased to the technical imaging specifications and patient population of the training data. In fact, only 6% of AI studies report external validation.[80]

2.7 Automated tools are needed for glioma

Gliomas are common infiltrative neoplasms of the central nervous system that affect patients of all ages, with variable growth rates and prognosis.[81,82] They are subdivided into four World Health Organization (WHO) grades (I-IV), based on the degree of differentiation, anaplasia, and aggressiveness.[83] WHO grades I and II are referred to as “low-grade” while WHO grades III and IV are considered “high-grade”. WHO grade I gliomas are considered the least malignant, while WHO grade II tumors (diffuse astrocytomas, oligodendrogliomas, and oligoastrocytomas) are more differentiated, and invariably progress to high grade.[84] WHO grades III (anaplastic astrocytomas, oligodendrogliomas, and oligoastrocytomas) and IV (glioblastoma and its variants) tumors are the most malignant, with a tendency to infiltrate into the surrounding brain parenchyma.

Glioblastoma (GBM) is the most common malignant primary adult brain tumor with five-year survival rates are less than 10%.[85,86] Despite active research in the treatment of GBM, the improvement of patient outcomes has lagged behind other types of cancers (Fig. 4-3).[87] The current standard of care is maximal safe surgical resection, chemoradiation, and adjuvant temozolomide. Within the natural history of GBM, there is inevitable recurrence of tumor, leading to patient death. There is currently no consensus on therapy for recurrent tumor as none have been proven to provide substantial survival benefit.[88] The current treatment strategy for GBM is suboptimal because clinicians do not have a reliable method of assessing tumor treatment response. An automatic tool that can longitudinally assess tumor volumes and genetic characteristics of the tumor would substantially improve evaluation of treatment efficacy, allowing for an earlier switch to alternative treatment strategies and thus, more personalized tailoring of patient care. To our knowledge, such a predictive tool for clinicians currently does not exist.

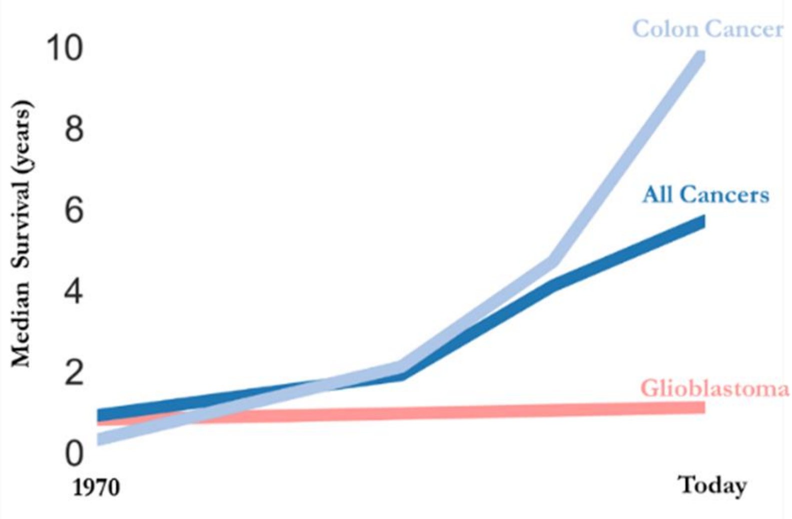


Figure 2-3. Median survival of colon cancer, all cancers, and glioblastoma over the last 4 decades.

The current clinical standard to measure tumor treatment response is based on the 2010 Response Assessment in Neuro-Oncology (RANO) criteria, which uses the product of maximal cross-sectional tumor diameters within T1 gadolinium post-contrast as measures of tumor burden.[89] However, the RANO criteria have several weaknesses: 1) It is manual and thus, time-consuming, 2) It is a two-dimensional measure (and thus, only a surrogate measure of tumor volume) 3) It is subject to inter- and intra-rater variability, and 4) It is sensitive to head position during imaging.[90–93] For these reasons, there has been high interest in developing a reproducible and accurate method for assessing tumor volumes. While manual volumetric segmentation of tumors is possible, it is difficult if the tumor is diffuse or demonstrates poor gadolinium contrast enhancement. Furthermore, the radiographic appearance of glioblastoma is quite heterogeneous, which makes delineation of boundaries challenging for even the most experienced neuroradiologists. As a result, manual segmentation is labor-intensive task and subject to variability, resulting in low reproducibility.[94,95] As such, there have been some automatic, deep learning methods proposed to automatically segment tumors pre-operatively.[96,97] There are currently no methods to that have been successfully validated for longitudinal segmentation of post-operative tumors and measurement of tumor burden.[98,99] Post-operative tumors are particularly challenging due to the presence of a resection cavity as well as blood products resulting from surgery. Because surgical resection is the standard of care for GBM, a tool for automatic segmentation in the post-operative setting is critically needed to allow for evaluation of treatment response.

In addition to assessing changes in tumor burden, evaluation of the underlying genetic characteristics of the tumor is needed to understand treatment efficacy. The most important molecular marker of gliomas is the presence of isocitrate dehydrogenase (*IDH)* mutations. In 2008, the presence of *IDH1* mutations, specifically involving the amino acid arginine at position 132, was demonstrated in in 12% of glioblastomas,[100] with subsequent reports observing *IDH1* mutations in 50-80% of LGGs.[101] In the wild-type form, the *IDH* gene product converts isocitrate into α-ketoglutarate [102]. When *IDH* is mutated, the conversion of isocitrate is instead driven to 2-

hydroxyglutarate, which inhibits downstream histone demethylases.[103] The presence of an *IDH* mutation carries important diagnostic and prognostic value. Gliomas with the *IDH1* mutation (or its homolog *IDH2*) carry a significantly increased overall survival than *IDH1/2* wild-type tumors, independent of histological grade.[100,104–106] Conversely, most lower grade gliomas with wild type IDH were molecularly and clinically similar to glioblastoma with equally dismal survival outcomes.[83] *IDH* wild-type grade III gliomas may in fact exhibit a worse prognosis than *IDH* mutant grade IV gliomas.[104] Its critical role in determining prognosis was emphasized with the inclusion of *IDH* mutation status as a classification parameter used in the 2016 update of WHO diagnostic criteria for gliomas.[107] Pre-treatment identification of *IDH* status can help guide clinical decision making. First, a priori knowledge of *IDH1* status with radiographic suspicion of a low-grade glioma may favor early intervention as opposed to observation as a management option. Second, *IDH* mutant gliomas are driven by specific epigenetic alterations, making them susceptible to therapeutic interventions (such as temozolomide) that are less effective against *IDH* wild-type tumors.[108,109] This is supported by *in vitro* experiments, which have found *IDH*-mutated cancer cells to have increased radio- and chemo-sensitivity.[110–112] Lastly, resection of non-enhancing tumor volume, beyond gross total removal of the enhancing tumor volume, was associated with a survival benefit in *IDH1* mutant grade III-IV gliomas but not in *IDH1* wild-type high-grade gliomas [113]. Thus, early determination of *IDH* status may guide surgical treatment plans, peri-operative counseling, and the choice of adjuvant management plans.

2.8 Automated tools are needed for ischemic stroke

Cerebrovascular disease is the third leading cause of death around the world after heart disease and cancer.[13,14] The most common clinical manifestation of cerebrovascular disease is an acute stroke, 87% of which are of an ischemic nature.[14] Symptoms often include focal neurological deficit, which can develop into chronic disease and disability. In fact, stroke is the leading cause of long-term disability worldwide in adults.[114] In the United States, the estimated direct and indirect costs of stroke is $68.9 billion.[115] Brain imaging is a key step in the clinical evaluation of ischemic stroke, with Computed Tomography (CT) and Diffusion Weighted Magnetic Resonance (DWI) being the key imaging modalities. While CT is more widely used due to its lower cost and acquisition time, DWI provides the advantage of being more sensitive.[116] Rapid and accurate evaluation is needed as intravenous thrombolysis should be performed within 4.5 hours of stroke onset.[117]

Important decisions in stroke management currently rely on accurate delineation of regions of acute ischemic brain injury. However, manual delineation of stroke regions is expensive, time-consuming, and subject to inter-rater variability. Furthermore, segmentation is a highly difficult task as there can be variability in size and location as well as ill-defined boundaries. As such, there has been efforts to develop automatic methods of performing lesion segmentation. Existing methods are limited in that they either require additional imaging modalities (such as T1-weighted, T2-weighted, and fluid attenuation inversion recovery, FLAIR) or they are semi-automatic and thus require manual input.[118–120] Recently, Chen et al. developed a fully automatic method using convolutional neural networks with the limitation that their approach only utilizes 2 dimensional information.[121]

Furthermore, there is a need to identify novel imaging subtypes that are predictive of stroke recovery. There are several distinct clinical sequelae of ischemic stroke, specifically patients with hyperglycemia, elevated troponin, prolonged QT, pneumonia, and urinary tract infections, which we hypothesize to be associated with different clinical subtypes. Two-thirds of stroke patients experience hyperglycemia, which is associated with adverse tissue outcomes.[122] Cardiac mortality accounts for 20% of stoke deaths.[123] Pneumonia and urinary tract infections occur in 7%- 22% and 3-40% of stroke patients, respectively.[124,125] The association between these clinical sequelae and anatomical imaging has been largely unexplored and has the potential to elucidate the biological mechanisms behind the clinical presentation of stroke. Using these imaging associations along with clinical covariates could be used to develop a useful tool to predict stroke recovery. Prediction of stroke recovery is pertinent as beta-blockers therapy has been shown to reduce mortality.[126] Furthermore, identification of patients of patients with poor predicted recovery can stratify patients who would benefit from supplemental treatment and management.

3 Distributed training of neural networks as an alternative to sharing patient data

3.1 Introduction

With the advent of powerful graphics processing units, deep learning has brought about major breakthroughs in tasks such as image classification, speech recognition, and natural language processing.[1–3] Due to the proficiency of neural networks at pattern recognition tasks, deep learning has created practical solutions to the challenging problem of automated support for clinical diagnosis. Recent studies have shown the potential of deep learning in detecting diabetic retinopathy, classifying dermatological lesions, predicting mutations in glioma, and assessing medical records.[12,40,127,128] Deep learning models take raw data as input and apply many layers of transformations to calculate a classification label of interest. The high dimensionality of these transformations allows these algorithms to learn complex patterns with a high level of abstraction.[4]

A requirement for the application of deep learning within the medical domain is a large quantity of training data, especially when the difference between imaging phenotypes is subtle or if there is large heterogeneity within the population. However, patient sample sizes are often small, especially for rarer diseases.[129] Small sample sizes may result in a neural network model with low generalizability.

A possible solution to the foregoing challenges is to perform a multicenter study, which can significantly increase the sample size as well as sample diversity. Ideally, patient data is shared to a central location where the algorithm can then be trained on all the patient data. However, there are challenges to this approach. First, if the patient data takes up a large amount of storage space (such as very high-resolution images), it may be cumbersome to share these data. Second, patient data is valuable, so institutions might simply prefer not to share data.[130] Third, there are often legal or ethical barriers to sharing patient data, making dispersal of some or all of the data not possible.[129]

In such cases, instead of sharing patient data directly, distributing the trained deep learning model may be a more appealing alternative. The model itself has much lower storage requirements than the patient data and does not contain any individually-identifiable patient information. Thus, distribution of deep learning models across institutions can overcome the limitations of distributing the patient data. However, the optimal method of performing such a task has not yet, to our knowledge, been studied.

One critical hurdle that prevents the deployment of deep learning models in the clinical work environment is their relatively poor generalizability across institutional differences, such as patient demographics, disease prevalence, scanners, and acquisition settings. A variety of recent deep learning studies that have shown poor generalizability of deep learning models when applied to data from different institutions than the one they were trained on.[78,79] Furthermore, the optimal method of distributing the process of training medical deep learning models across heterogenous institutions has not yet been adequately studied. Indeed, much of the medical deep learning studies have been on independent and identically distributed (IID) data.[5,131] In this scenario, the institutions have no intra-institution correlation and the data across institutions is identically distributed.[132] More work needs to be done on dealing with dataset skew (non-IID data) across institutions, specifically when there is: 1) quantity skew (e.g. a large academic hospital has significantly more data than a small community hospital), 2) feature distribution skew (e.g. one hospital uses one scanner vendor while another hospital uses a different scanner vendor), 3) label distribution skew (e.g. obesity is much more prevalent in North American than in Asia), 4) concept shift – same label, different features (e.g. eczema looks different on light vs dark skin), and 5) concept shift- same features, different labels (e.g. physicians in North America may be more conservative in calling a certain disease than physicians in Asia due to higher rates of litigation for unnecessary treatment).[132,133] In a real case scenario, the data across institutions will contain a mixture of skew types, which makes the problem even more challenging (Fig. 3-1).

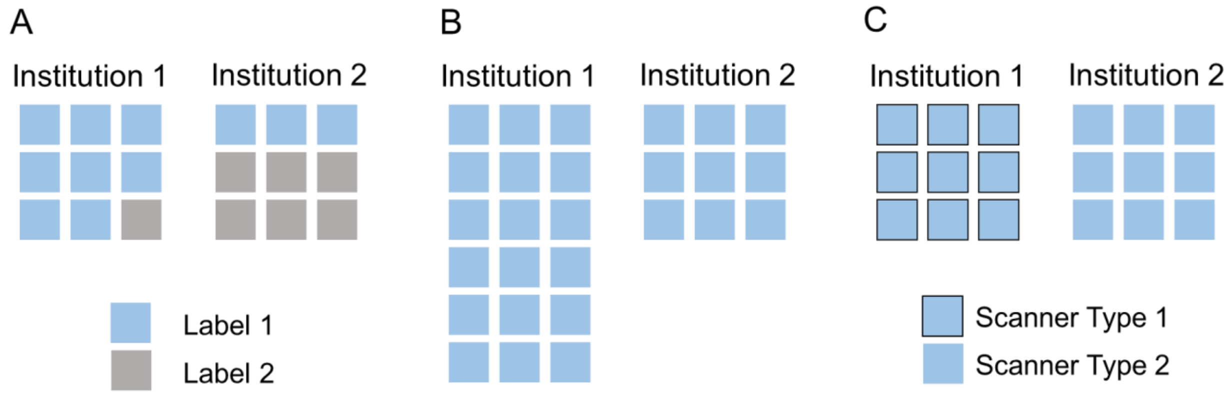


Figure 3-1. Various types of heterogeneity can exist in real patient data, such as (A) imbalanced labels or patient characteristics (label distribution skew or concept shift), (B) data size heterogeneity (quantity skew), and (C) differences in data acquisition (feature distribution skew)

In this section, we simulate the distribution of deep learning models across institutions using various heuristics. We compare the results with a deep learning model trained on centrally hosted patient data. We demonstrate these simulations on 3 datasets: retinal fundus photos, mammography, and ImageNet. We aim to assess 1) the performance of distributing deep learning models compared to sharing patient data, 2) whether the performance distributing deep learning models is compromised when variability is introduced to an institution, and 3) if distributing deep learning models can achieve high performance on a large scale (that is, when there are many institutions).

3.2 Methods

3.2.1 Preprocessing (initial image collection)

We obtained 35,126 color digital retinal fundus (interior surface of the eye) photos from the Kaggle Diabetic Retinopathy competition[134]. Each image was rated for disease severity by a licensed clinician on a scale of 0-4 (absent, mild, moderate, severe, and proliferative retinopathy, respectively). The images came from 17,563 patients of multiple primary care sites throughout California and elsewhere. The acquisition conditions were varied, with a range of camera models, levels of focus, and exposures. In addition, the resolutions ranged from 433x289 pixels to 5184x3456 pixels[135]. The images were pre-processed via the method detailed in the competition report by the winner, Ben Graham[136]. To summarize his method, the OpenCV python package was used to rescale images to a radius of 300, followed by local color averaging and image clipping. The images were then resized to 256x256 to reduce the memory requirements for training the neural network. To simplify training of the network, the labels were binarized to Healthy (scale 0) and Diseased (scale 2, 3, or 4). Furthermore, mild diabetic retinopathy images (scale 1, n = 2443 images), which represent a middle ground between Healthy and Diseased, were not used for our experiments. It is also known that there is a correlation between the disease status of the left eye and the status of the right eye. To remove this as a confounding factor in our study, only images from left eye were utilized.

3.2.2 Convolutional neural network

We utilized the 34-layer residual network (ResNet34) architecture (Fig. 3-2A)[137]. Our implementation was based on the Keras package with Theano backend. [138,139] The convolutional neural networks were run on a NVIDIA Tesla P100 GPU. During training, the probability of samples belonging to Healthy or Diseased class was computed with a sigmoid classifier. The weights of the network were optimized via a stochastic gradient descent algorithm with a mini-batch size of 32. The objective function used was binary cross-entropy. The learning rate was set to .0005 and momentum coefficient of .9. The learning rate was multiplied by .25 when the same training images were used to train the neural network 20 times with no improvement of the validation loss. The learning rate was decayed a total of 3 times (Training Phases A-D, Fig 3-2C). Biases were initialized using the Glorot uniform initializer.[140] To prevent overfitting and to improve learning, we augmented the data in real-time by introducing random rotations (0-360 degrees) and flips (50% change of horizontal or vertical) of the images at every epoch. The final model was evaluated by calculating the accuracy on the unseen testing cohort.

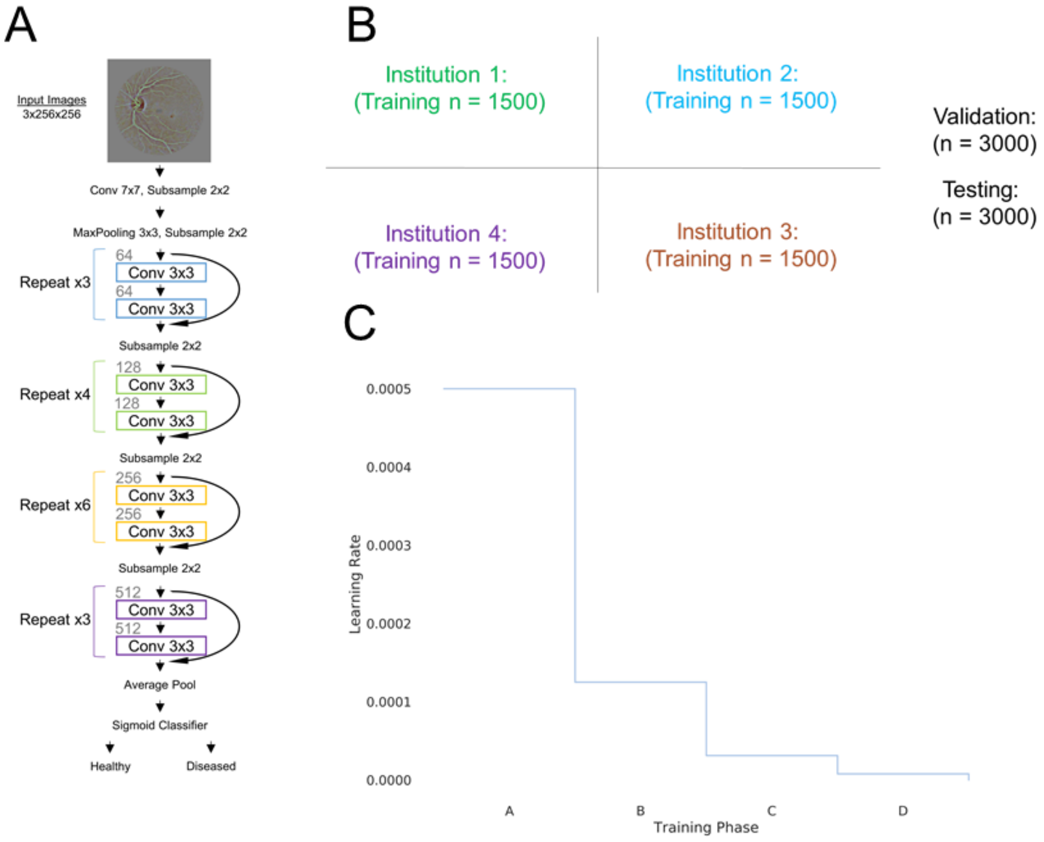


Figure 3-2. (A) ResNet-34 architecture was utilized for the Diabetic Retinopathy dataset. (B) The dataset was randomly divided into 4 institutions along with a validation and testing set. (C) The learning rate was decayed to .25 of its value when the same input samples are inputted into the network 20 times at a given learning rate without an improvement of the validation loss.[5]

3.2.3 Model training heuristics with 4 institutions

The dataset was randomly sampled, with equal class distributions, into 4 “institutions”, each institution having n = 1500 patients. In addition, the dataset was sampled to create a single validation cohort (n = 3000 patients) and a single testing (n = 3000 patients) cohort, again with equal class probabilities (Fig. 3-2B). Sampling was without replacement such that there are no overlapping patients in any of the cohorts. The image intensity was normalized within each channel across all patients within each cohort. Because model performance plateaus as the number of training patient samples increases, the number of patients per institution was limited to 1500 to prevent saturation of learning for models trained in single institutions.

We tested several different training heuristics (Fig. 3-3) and compared the results. The first heuristic is training a neural network for each institution individually, assuming there is no collaboration between the institutions. The second heuristic is collaboration through pooling of all patient data into a shared dataset (centrally hosted data, Fig. 3-3A). The third heuristic was averaging the output of the four models trained on the institutions individually (ensemble single institution models, Fig. 3-3B). For n participating institutions, I_1, I_2, …, and I_n, the output, O, would be: $O = \frac{1}{n}\sum_{i=1}^{n} O_i$. Ensembling is typically used to gain some performance over a single model due to the stochastic nature of these models.[141] The fourth heuristic was training a model at a single institution until plateau of validation loss and then transferring the model to the next institution (single weight transfer, Fig. 3-3C). Under the single weight transfer training heuristic, the model is transferred to each institution exactly once. This is akin to transfer learning, whereby a model is trained on one dataset and the fine-tuned on another.[142] The last heuristic was training a model at each institution for a predetermined number of epochs (weight transfer frequency) before transferring the model to the next institution (cyclical weight transfer, Fig. 3-3D). The idea of cyclical weight transfer is that the model sees all the data till convergence similar to when the model is trained on centrally located data. Under the cyclical weight transfer training heuristic, the model is transferred to each institution more than once. The frequencies of weight transfer we studied were every 20 epochs, 10 epochs, 5 epochs, 4 epochs, 2 epochs, and every epoch.

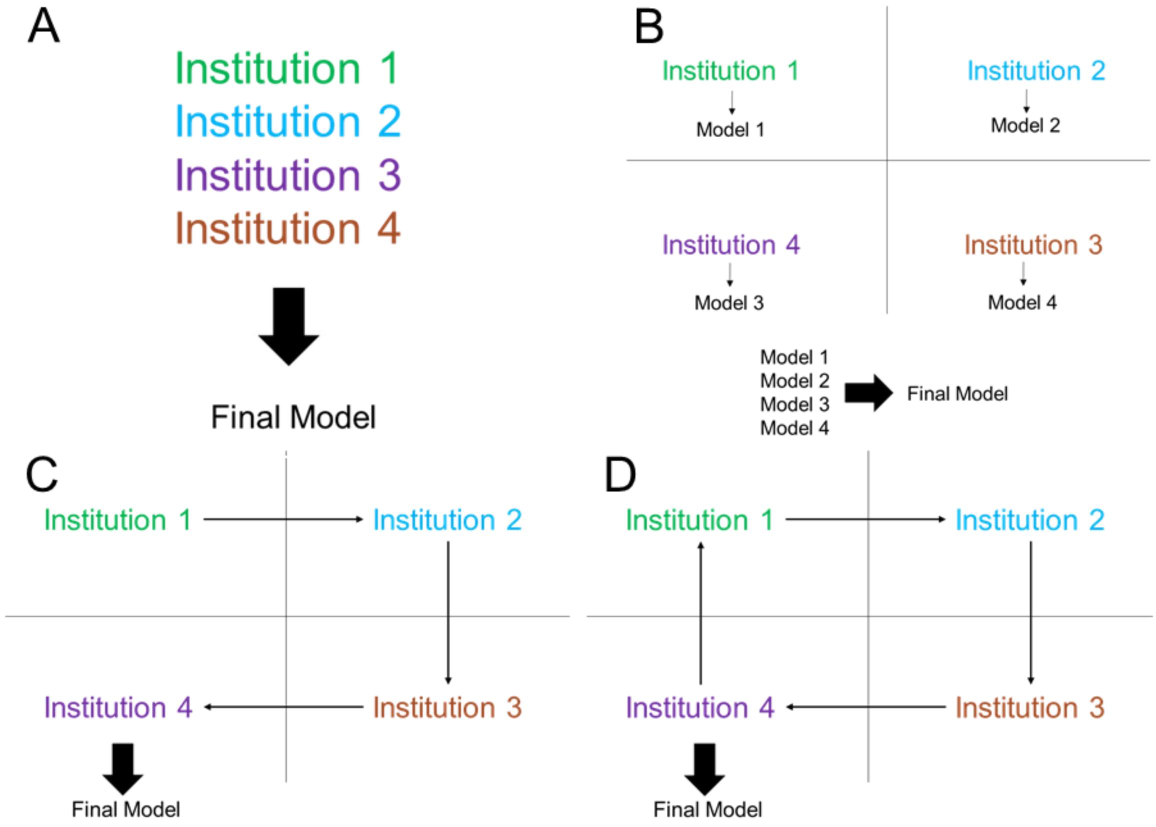


Figure 3-3. Model training heuristics investigated include (A) centrally hosted, (B) ensemble single institution models, (C) single weight transfer, and (D) cyclical weight transfer.[5]

3.2.4 Cyclical weight transfer with 20 institutions

We next addressed whether cyclical weight transfer can improve model performance when the performance of any individual institution is no better than random classification. This simulates a scenario where data at any individual institution is sparse such as in a community hospital or for a rare disease. To do this, we divided 6000 patient samples from the Kaggle Diabetic Retinopathy dataset into 20 institutions (n = 300 per institution) with equal class distributions. As with our previous experiments, we also sampled a single validation cohort (n = 3000 patient samples) and a single testing cohort (n = 3000 patient samples) with equal class probabilities. We then performed experiments with different numbers of collaborating institutions, starting with 1 and increasing to all 20 institutions. We utilized the cyclical weight transfer training heuristic with a weight transfer frequency of 1 epoch. We evaluated model performance via testing cohort accuracy. We compared testing accuracies with that of random classification and with the testing accuracy of a model trained with all 6000 patient samples centrally hosted.

3.2.5 Introduction of an institution with variability

In our initial division of the different institutions, we assumed that each institution had the same number of patients, ratio of healthy to diseased patients, and image quality. However, in a real scenario, there will likely be variability within institutions that may compromise the predictive performance of the model. These modes of variability include differences in patient demographics, disease prevalence, and image acquisition settings (image resolution, detector sensitivity, heterogeneity in spatial sensitivity, data post-processing, and modality specific parameters such as field strength, echo time, and repetition time for MR).

To simulate this possibility, we introduced variability into one of the 4 institutions and assessed the performance of the different training heuristics. We simulated two scenarios: In the first, we decreased the resolution of the images by a factor of 16. In the second, we significantly decreased the number of patients (from n = 1500 to n = 150) and introduced class imbalance (ratio of healthy to diseased was 9:1). We assessed the performance of centrally hosted data, ensembling single institution models, single weight transfer, and cyclical weight transfer with weight transfer at every epoch. For single weight transfer, we experimented with ordering of the institutions, specifically whether the variable institution was Institution 1, 2, 3, or 4. For cyclical weight transfer, we assessed the performance of not skipping vs skipping the variable institution entirely. A summary of all experiments performed with the Kaggle Retinopathy Dataset is summarized in Table 3-1.

Experiment	Summary
Model Training Heuristics with 4 Institutions	In this experiment, there are 4 equivalent institutions. We evaluate the performance of model ensembling, single weight transfer, and cyclical weight transfer compared to centrally hosted patient data.
Cyclical Weight Transfer with 20 Institutions	In this experiment, there are 20 institutions. The number of patients at each institution is such that a model trained on patients from a single institution is no better than random classification. We evaluate the performance of cyclical weight transfer as the number of collaborating institutions increase from 1 to all 20.
Introduction of an Institution with Variability	In this experiment, there are 4 institutions but one of the institutions has a mode of variability introduced (either low-resolution images or a low number of patients with class imbalance). We evaluate the effectiveness of model ensembling, single weight transfer, and cyclical weight transfer compared to centrally hosted patient data.

Table 3-1. A summary of all experiments performed with the Kaggle Retinopathy Dataset.

3.2.6 Repetition of experiment in a second image collection

To demonstrate the reproducibility of our results, we repeated our experiment on model training heuristics with 4 institutions on the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (DDSM) dataset, an open source labeled dataset of mammograms[143]. For each patient, the dataset includes cranial-caudal and/or mediolateral-oblique views of the right and/or left breast, and each image is labeled as benign or malignant. For our experiments, we use a subset of 1508 grayscale images from 800 patients that had a mass in the breast. Along with each image, a binary segmentation mask for the mass was available. Of the 1508 images, 722 were labeled malignant and 786 were labeled benign, so a majority classifier would have 52.1% accuracy. We randomly selected 140 patients for each of the 4 “institutions”, 120 patients for the validation cohort, and 120 patients for the testing cohort. This resulted in 257 images in Institution 1, 266 images in Institution 2, 257 images in Institution 3, and 270 images in Institution 4 (total of 1050 training images), 229 images in the validation set, and 229 images in the testing set. For the same patient, the different images, including different views of the same breast, could have different labels. Thus, we treated each image separately, but did not allow images from the same patient to be divided across different institutions, or across the training and testing/validation cohorts (as in our experiments with the Kaggle Diabetic Retinopathy dataset).

The grayscale image pixels were scaled between 0 and 1, and the mask pixels were either 0 or 1. Each image was cropped into 256x256 pixel resolution such that the region of interest as indicated by the binary mask was centered in the largest possible bounding box. Each cropped grayscale image along with its corresponding cropped binary mask were combined to produce a 2-channel 256x256 image. The images were normalized by subtracting the maximum pixel intensity and zero-centered by subtracting the mean pixel intensity. These normalized 2x256x256 images were input into a neural network. For this dataset, we used a 22-layer GoogLeNet with batch normalization after each convolutional layer, batch size of 32, and dropout of 0.5 before the final readout layer.[144] We used Adam optimizer with initial learning rate of 0.001 and learning rate decay of 0.99 every epoch (every 4 epochs in the weight transfer experiments) to optimize the model. [145] Cross entropy with L2 regularization coefficient of 0.0001 was used as the loss function. Model learning is terminated when there were 80 epochs of no improvement in validation loss (320 epochs in the weight transfer experiments). For the single weight transfer experiment, weights were transferred to the next institution each time there were 20 epochs of no improvement in validation loss, and learning was terminated when there were 20 epochs having no improvement in validation loss at the final institution. For ensembling, the output probabilities from the models trained at each of the 4 institutions were averaged to produce final class predictions. During training, the data were augmented by introducing random rotations (0-360 degrees) and flips (50% change of horizontal or vertical) to the images at every epoch.

3.2.7 Repetition of experiment in a non-medical image collection

We further demonstrate the reproducibility of our results by repeating our experiment on model training heuristics with 4 institutions on the ImageNet dataset[16]. We utilized the ImageNet 2012 classification dataset, which contains 1.28 million training images and 1000 classes. To decrease the time of training, we utilized a subset of the training images for our experiments. We randomly selected 20 classes of the 1000 to work with. We randomly allocated 75 images of each class to each "institution" and 150 images of each class to the validation and testing cohort. In total, each of the 4 institutions had 1500 images and both the validation and testing cohorts had 3000 images. For pre-processing, we resized each image to 224x224 and subtracted by the per channel mean of the entire ImageNet dataset. As with the experiments with the Kaggle Diabetic Retinopathy dataset, we utilized the 34-layer residual network architecture. The learning rate was set to .0001 and momentum coefficient was set to .9. The learning rate was decayed to .25 of its value when the same samples were inputted into the network 20 times at a given learning rate with no improvement of the validation loss. To prevent overfitting and improve learning, we augmented the data by introducing random rotations (0-360 degrees), flips (50% change of horizontal or vertical), zooming (from -20% to +20%), and shearing (0 to .2 radians) at every epoch. We evaluated our models by assessing both the top-1 and top-5 accuracies. Top-1 accuracy is calculated by comparing the ground truth label with the top predicted class. Top-5 accuracy is calculated by comparing the ground truth label with the top 5 predicted classes.

3.3 Results

3.3.1 Retinal fundus dataset - single institution training

The models trained on single institutions had poor performance (Fig. 3-4A-D) due to dataset size limitations. The average testing accuracies for the single institution models was 56.3% (Table 3-2). The highest testing accuracy for a network trained on a single institution was 59.0%.

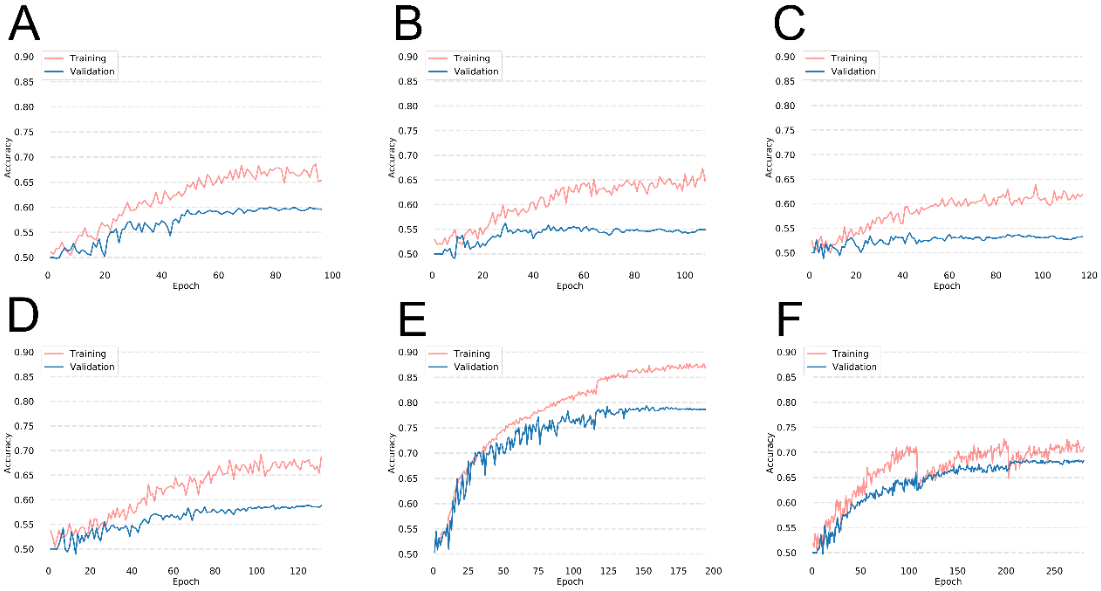


Figure 3-4. Performance of a neural network when trained on (A) Institution 1, (B) Institution 2, (C) Institution 3, and (D) Institution 4 for the Diabetic Retinopathy dataset. The training and validation accuracies for a model trained the centrally hosted training and single weight transfer training heuristics are shown in (E) and (F), respectively.[5]

<table>
<tr><th>Diabetic Retinopathy</th><th colspan="2">Training Accuracy
(n = 1500, %)</th><th colspan="2">Validation Accuracy
(n = 3000, %)</th><th colspan="2">Testing Accuracy
(n = 3000, %)</th></tr>
<tr><td>Institution 1</td><td colspan="2">68.1</td><td colspan="2">59.6</td><td colspan="2">59.0</td></tr>
<tr><td>Institution 2</td><td colspan="2">66.8</td><td colspan="2">54.9</td><td colspan="2">53.8</td></tr>
<tr><td>Institution 3</td><td colspan="2">64.3</td><td colspan="2">53.3</td><td colspan="2">54.3</td></tr>
<tr><td>Institution 4</td><td colspan="2">69.5</td><td colspan="2">58.8</td><td colspan="2">58.2</td></tr>
<tr><th>DDSM</th><th colspan="2">Training Accuracy
(n = 257-270, %)</th><th colspan="2">Validation Accuracy
(n = 229, %)</th><th colspan="2">Testing Accuracy
(n = 229, %)</th></tr>
<tr><td>Institution 1</td><td colspan="2">59.1</td><td colspan="2">55.5</td><td colspan="2">55.0</td></tr>
<tr><td>Institution 2</td><td colspan="2">56.1</td><td colspan="2">57.2</td><td colspan="2">52.8</td></tr>
<tr><td>Institution 3</td><td colspan="2">59.0</td><td colspan="2">52.8</td><td colspan="2">60.3</td></tr>
<tr><td>Institution 4</td><td colspan="2">61.6</td><td colspan="2">56.3</td><td colspan="2">54.6</td></tr>
<tr><th>ImageNet</th><th colspan="2">Training Accuracy
(n = 1500, %)</th><th colspan="2">Validation Accuracy
(n = 3000, %)</th><th colspan="2">Testing Accuracy
(n = 3000, %)</th></tr>
<tr><td></td><td>Top-1</td><td>Top-5</td><td>Top-1</td><td>Top-5</td><td>Top-1</td><td>Top-5</td></tr>
<tr><td>Institution 1</td><td>62.1</td><td>93.5</td><td>30.4</td><td>71.4</td><td>31.0</td><td>71.2</td></tr>
<tr><td>Institution 2</td><td>66.1</td><td>95.0</td><td>31.1</td><td>70.0</td><td>32.4</td><td>71.5</td></tr>
<tr><td>Institution 3</td><td>64.5</td><td>94.3</td><td>31.5</td><td>71.3</td><td>32.4</td><td>71.1</td></tr>
<tr><td>Institution 4</td><td>66.8</td><td>94.5</td><td>31.6</td><td>70.8</td><td>32.1</td><td>71.6</td></tr>
</table>

Table 3-2. Training, validation, and testing accuracy of the neural network when trained on single institutions for the Diabetic Retinopathy, DDSM, and ImageNet datasets.

3.3.2 Centrally hosted training

When patient data from all institutions were pooled together, the collective size of the dataset was 6000. A network trained on the combined dataset had a high performance with a testing accuracy of 78.7% (Table 3-3).

<table>
<tr><th>Diabetic Retinopathy</th><th colspan="2">Training Accuracy
(n = 6000, %)</th><th colspan="2">Validation Accuracy
(n = 3000, %)</th><th colspan="2">Testing Accuracy
(n = 3000, %)</th></tr>
<tr><td>Centrally Hosted</td><td colspan="2">89.4</td><td colspan="2">78.6</td><td colspan="2">78.7</td></tr>
<tr><td>Ensemble Models</td><td colspan="2">63.2</td><td colspan="2">60.9</td><td colspan="2">60.0</td></tr>
<tr><td>Single Weight Transfer</td><td colspan="2">70.4</td><td colspan="2">68.3</td><td colspan="2">68.1</td></tr>
<tr><td>DDSM</td><td colspan="2">Training Accuracy
(n = 1050, %)</td><td colspan="2">Validation Accuracy
(n = 229, %)</td><td colspan="2">Testing Accuracy
(n = 229, %)</td></tr>
<tr><td>Centrally Hosted</td><td colspan="2">77.0</td><td colspan="2">71.6</td><td colspan="2">70.7</td></tr>
<tr><td>Ensemble Models</td><td colspan="2">63.7</td><td colspan="2">56.3</td><td colspan="2">61.1</td></tr>
<tr><td>Single Weight Transfer</td><td colspan="2">61.3 ± 0.9</td><td colspan="2">61.2 ± 0.8</td><td colspan="2">61.1 ± 1.8</td></tr>
<tr><td>ImageNet</td><td colspan="2">Training Accuracy
(n = 6000, %)</td><td colspan="2">Validation Accuracy
(n = 3000, %)</td><td colspan="2">Testing Accuracy
(n = 3000, %)</td></tr>
<tr><td></td><td>Top-1</td><td>Top-5</td><td>Top-1</td><td>Top-5</td><td>Top-1</td><td>Top-5</td></tr>
<tr><td>Centrally Hosted</td><td>82.9</td><td>98.4</td><td>49.5</td><td>83.4</td><td>48.9</td><td>83.8</td></tr>
<tr><td>Ensemble Models</td><td>50.2</td><td>88.6</td><td>37.0</td><td>76.5</td><td>38.6</td><td>77.0</td></tr>
<tr><td>Single Weight Transfer</td><td>45.5</td><td>84.5</td><td>36.0</td><td>76.2</td><td>37.9</td><td>75.5</td></tr>
</table>

Table 3-3. Training, validation, and testing accuracy of centrally hosted training, ensembling single institution model outputs, and single weight transfer for for Diabetic Retinopathy, DDSM, and ImageNet datasets.

3.3.3. Ensembling single institution models

Averaging the sigmoid probability of the single institution models resulted in a testing accuracy of 60.0% (Table 3-3). Notably, the ensembled model outperformed any network trained on a single institution in terms of validation and testing accuracy.

3.3.4 Single weight transfer

Using single weight transfer heuristic, the model was trained at each institution until the plateau of validation loss was reached, followed by transferring of the model to the next institution. The resulting model had a testing accuracy of 68.1% (Table 3-3).

3.3.5 Cyclical weight transfer

In our initial experiment, we trained the network for 20 epochs at each institution before transferring the weights to the next institution. The average testing accuracy after repeating this experiment 3 times was 76.1% (Table 3-4).

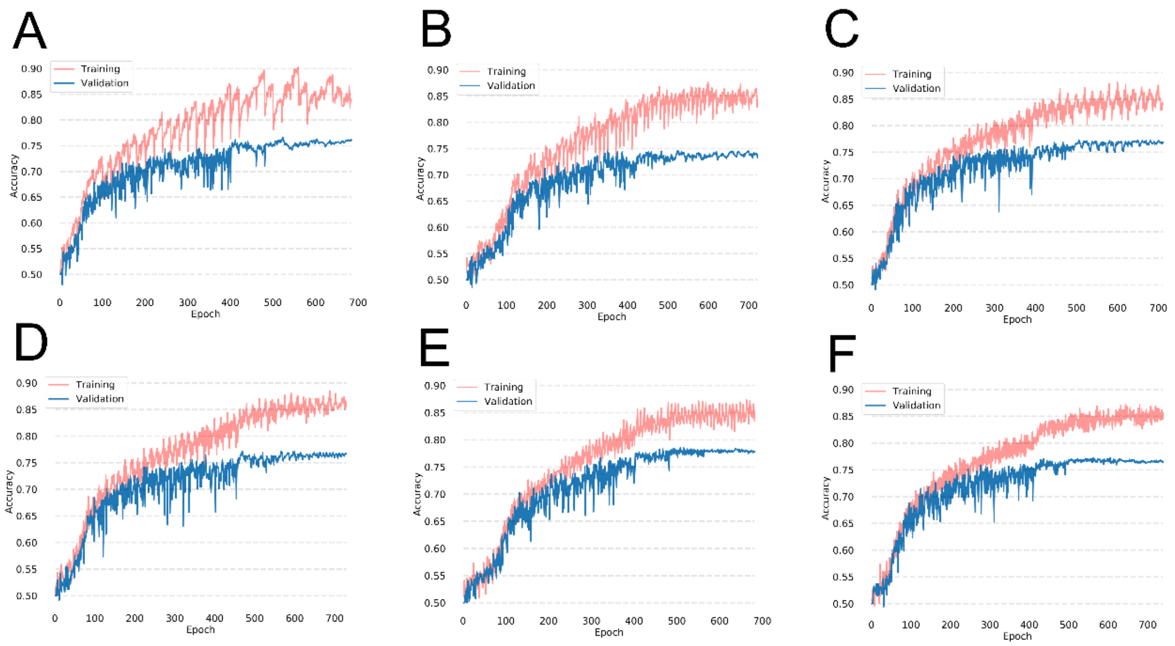


Figure 3-5. Training and validation accuracies during training on the Diabetic Retinopathy dataset with cyclical weight transfer with weight transfer frequencies of every (A) 20 epochs, (B) 10 epochs, (C) 5 epochs, (D) 4 epochs, (E) 2 epochs, or (F) every epoch.[5]

<table>
<tr><td>Diabetic Retinopathy</td><td colspan="2">Training Accuracy
(n = 6000, %)</td><td colspan="2">Validation Accuracy
(n = 3000, %)</td><td colspan="2">Testing Accuracy
(n = 3000, %)</td></tr>
<tr><td>Cyclical Weight Transfer, Every:
20 Epochs
10 Epochs
5 Epochs
4 Epochs
2 Epochs
Epoch</td><td colspan="2">85.8 ± 0.9
87.9 ± 1.6
86.8 ± 0.9
88.9 ± 1.1
89.1 ± 1.7
89.4 ± 2.3</td><td colspan="2">76.0 ± 0.6
75.6 ± 2.0
76.1 ± 0.6
76.6 ± 0.1
77.3 ± 0.5
77.3 ± 1.3</td><td colspan="2">76.1 ± 1.0
75.9 ± 1.2
76.1 ± 0.8
77.4 ± 0.2
77.8 ± 0.3
77.3 ± 0.9</td></tr>
<tr><td>DDSM</td><td colspan="2">Training Accuracy
(n = 1050, %)</td><td colspan="2">Validation Accuracy
(n = 229, %)</td><td colspan="2">Testing Accuracy
(n = 229, %)</td></tr>
<tr><td>Cyclical Weight Transfer, Every:
20 Epochs
10 Epochs
5 Epochs
4 Epochs
2 Epochs
Epoch</td><td colspan="2">72.7 ± 1.3
70.5 ± 4.7
71.5 ± 3.0
71.7 ± 1.9
71.9 ± 1.5
74.8 ± 2.0</td><td colspan="2">66.5 ± 3.5
68.9 ± 0.9
69.1 ± 0.2
65.9 ± 1.8
69.3 ± 2.4
68.9 ± 1.3</td><td colspan="2">65.4 ± 1.1
68.1 ± 3.6
68.1 ± 1.2
68.7 ± 2.4
69.9 ± 2.7
69.1 ± 2.9</td></tr>
<tr><td>ImageNet</td><td colspan="2">Training Accuracy
(n = 6000, %)</td><td colspan="2">Validation Accuracy
(n = 3000, %)</td><td colspan="2">Testing Accuracy
(n = 3000, %)</td></tr>
<tr><td></td><td>Top-1</td><td>Top-5</td><td>Top-1</td><td>Top-5</td><td>Top-1</td><td>Top-5</td></tr>
</table>

Cyclical Weight Transfer, Every:						
20 Epochs	77.2 ± 3.2	97.7 ± 0.8	46.9 ± 0.8	82.8 ± 0.7	46.6 ± 0.9	83.2 ± 0.9
10 Epochs	78.5 ± 1.2	98.0 ± 0.4	47.8 ± 0.9	82.9 ± 0.4	47.3 ± 0.6	83.8 ± 0.1
5 Epochs	77.7 ± 2.6	97.7 ± 0.4	47.7 ± 0.7	83.0 ± 0.1	47.5 ± 1.4	83.3 ± 0.5
4 Epochs	78.5 ± 3.5	97.9 ± 0.6	47.2 ± 0.9	83.2 ± 0.5	48.1 ± 0.6	83.6 ± 0.2
2 Epochs	79.0 ± 3.2	97.8 ± 0.9	47.9 ± 0.0	82.8 ± 0.4	47.6 ± 1.1	84.1 ± 0.4
Epoch	83.2 ± 3.5	98.6 ± 0.6	49.2 ± 0.3	83.9 ± 0.7	49.3 ± 1.0	84.7 ± 0.1

Table 3-4. Training, validation, and testing accuracy for cyclical weight transfer for Diabetic Retinopathy, DDSM, and ImageNet datasets. Weight transfer frequencies investigated include every 20 epochs, 10 epochs, 5 epochs, 4 epochs, 2 epochs, and epoch. The accuracies for cyclical weight transfer are shown as mean ± standard deviation for 3 repetitions.

We also investigated whether having a higher frequency of weight transfer can improve the testing accuracy. We experimented with weight transfer frequencies of 10, 5, 4, 2, and every epoch, repeating each experiment 3 times (Table 3-4). The average testing accuracy of lower frequency weight transfer (every 20, 10, or 5 epochs) was 76.1% while the average testing accuracy of higher frequency weight transfer (every 4, 2, or 1 epoch) was 77.5% (two-sample t-test p-value < .001). Thus, a higher frequency weight transfer had a statistically significant increase in testing accuracy. The average training testing accuracy for all cyclical weight transfer experiments was 76.8%.

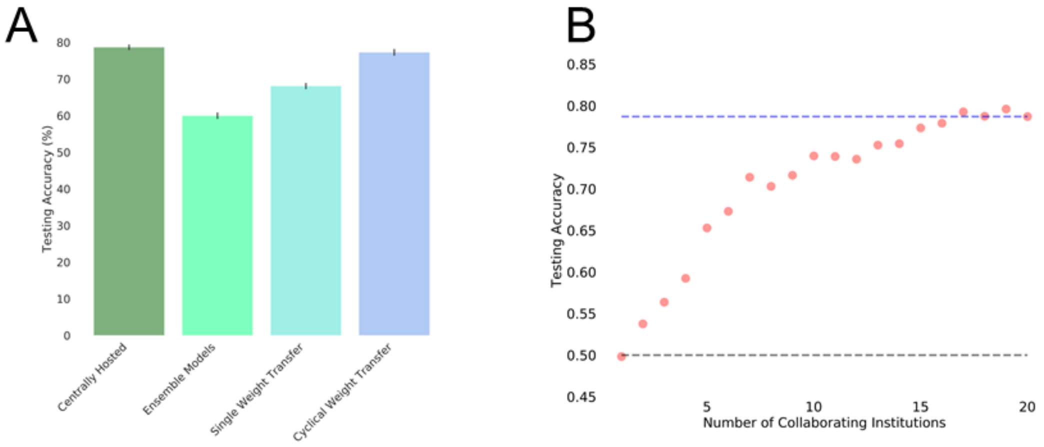


Figure 3-6. (A) Testing accuracies of centrally hosted training, ensembling models, single weight transfer, and cyclical weight transfer for our 4 “institution” experiment on the Diabetic Retinopathy dataset. Cyclical weight transfer had the performance that was on par with centrally hosted training ($p > .05$). (B) To show distributed computation on a larger scale, we performed a 20 “institution” experiment with n = 300 patients per institution. The plot shown is the testing accuracy as a function of the number of collaborating institutions. All models were trained using the cyclical weight transfer training heuristic with a weight exchange frequency of 1. For reference, testing accuracy expected from random classification (gray line) and centrally hosted data (n = 6000 patients, blue line) are shown.[5] When all 20 institutions participated in cyclical weight transfer, the performance was not different from that of centrally hosted data ($p > .05$).

3.3.6 Cyclical Weight Transfer With 20 Institutions

We next addressed whether cyclical weight transfer can improve model performance when the performance of any individual institution is no better than random classification. To do this, we divided 6000 patient samples into 20 institutions, each with n = 300 patients. We trained models with increasing numbers of collaborating institutions, from 1 to 20. We utilized the cyclical weight transfer training heuristic with the weight transfer frequency of 1. As we increased the number of collaborating institutions, the testing accuracy increased (Fig. 3-6B). The testing accuracy for a single institution was 49.8%, which is equivalent to random classification as there are equal numbers of healthy and diseased patients. The testing accuracy for 20 collaborating institutions was 78.7%, which is on par with the performance of centrally hosted data with all 6000 patient samples.

3.3.7 Introduction of an institution with variability

We next addressed what would happen if variability was introduced into one of the institutions. The modes of variability were either an institution with low-resolution images or an institution with few patients and class-imbalance. Among the various model-sharing training heuristics that was trained on all 4 institutions, cyclical weight transfer had the highest testing performance (Table 3-5), with a testing accuracy of 72.7% in experiments with an institution with low-resolution images and 73.3% in experiments with an institution with a small number of patients with class-imbalance. This is of comparable performance to that of centrally hosted data, which had testing accuracies of 72.2% and 75.4%, respectively. It is interesting to note that the performance of single weight transfer was dependent on the ordering of the institutions (that is, whether the variable institution was institution 1, 2, 3, or 4), which can be attributed to catastrophic forgetting.[146] We also assessed performance of cyclical weight transfer when the variable institution was skipped. The resulting testing accuracy was 74.4%, which is comparable to cyclical weight transfer that included the variable institution.

	Variable Institution: Low-Resolution	Variable Institution: Small and Imbalanced

	Testing Accuracy (n = 3000, %)	Testing Accuracy (n = 3000, %)
Centrally Hosted	72.2	75.4
Ensembling Models	57.8	58.9
Single Weight Transfer (Variable Institution as Institution 1)	55.2	54.7
Single Weight Transfer (Variable Institution as Institution 2)	64.6	67.6
Single Weight Transfer (Variable Institution as Institution 3)	57.4	67.2
Single Weight Transfer (Variable Institution as Institution 4)	50.4	64.3
Cyclical Weight Transfer, Every Epoch	72.7	73.3
Cyclical Weight Transfer, Every Epoch (Skipping Variable Institution)	74.4	

Table 3-5. The testing accuracy of the various training heuristics with the various training heuristics when variability (low-resolution images or few patients with class-imbalance) was introduced into one of the institutions.

3.3.8 Mammography dataset

When we repeated the experiments on the DDSM dataset, the average testing accuracy was 55.7% for single institution models (Table 3-2, Fig. 3-7A-D), only slightly better than a majority classifier. A model trained on centrally hosted data had a testing accuracy of 70.7% (Table 3-3, Fig. 3-7E). Ensembling single institution models (via averaging of the model outputs) resulted in a testing accuracy of 61.1% and the single weight transfer training heuristic also resulted in an average testing accuracy of 61.1% (Table 3-3, Fig. 3-7F). Cyclical weight transfer resulted in an average testing accuracy of 67.2% for low frequencies of weight transfer (every 20, 10, or 5 epochs), which was lower than the average testing accuracy of 69.2% for high frequency of weight transfer (every 4, 2, or 1 epoch, $p < .05$) (Fig. 3-8, Table 3-4).

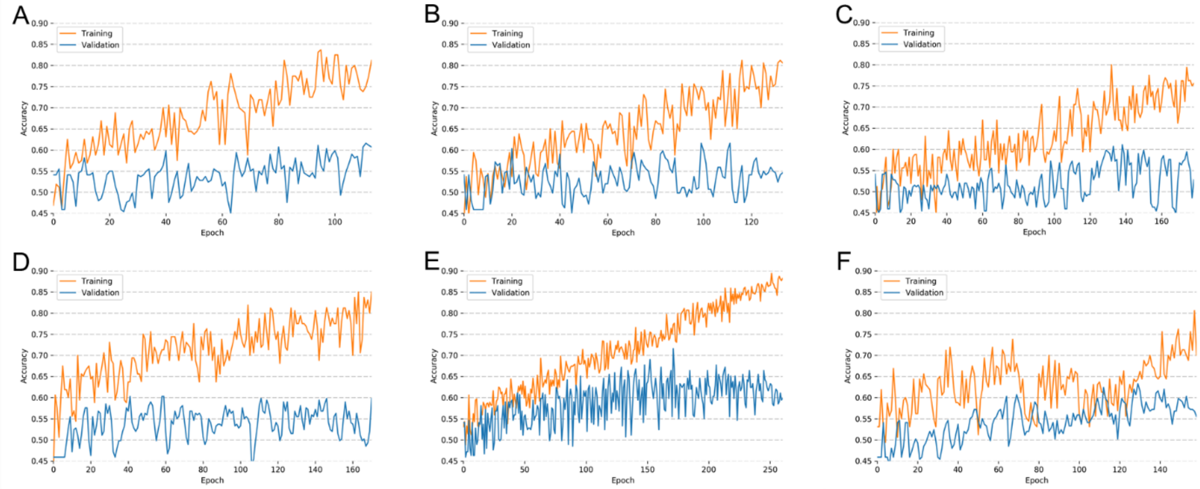


Figure 3-7. Training and validation accuracies during training on DDSM dataset when trained on (A) Institution 1, (B) Institution 2, (C) Institution 3, and (D) Institution 4, (E) Centrally Hosted Training Heuristic, and (F) Single Weight Transfer Training Heuristic.

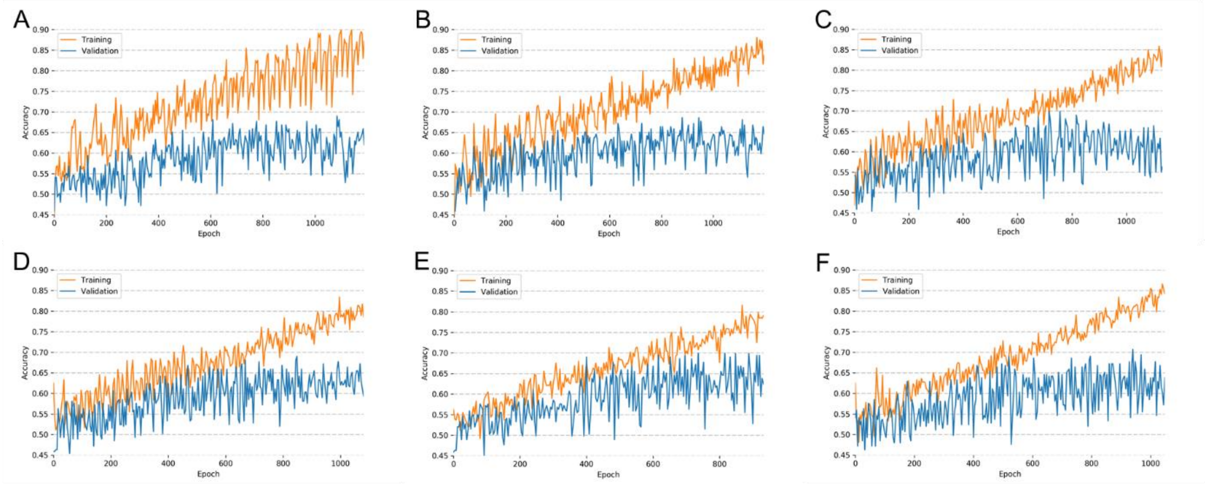


Figure 3-8. Training and validation accuracies during training on the DDSM dataset with cyclical weight transfer with weight transfer frequencies of every (A) 20 epochs, (B) 10 epochs, (C) 5 epochs, (D) 4 epochs, (E) 2 epochs, or (F) every epoch.

3.3.9 ImageNet dataset

When these experiments were repeated for the ImageNet dataset, the average testing top-1 accuracy was 32.0% (top-5 accuracy = 71.4%) for single institution models (Table 3-2, Fig. 3-9A-D). In comparison, a model trained on centrally hosted data had a testing top-1 accuracy of 48.9% (top-5 accuracy = 83.8%) (Table 3, Fig. 3=9E). Ensembling single institution models resulted in a testing top-1 accuracy of 38.6% (top-5 accuracy = 77.0%), while the single weight transfer training heuristic resulted in a testing top-1 accuracy of 37.9% (top-5 accuracy = 75.5%) (Table 3-3, Fig. 3-9F). Cyclical weight transfer resulted in an average testing top-1 accuracy of 47.1% (top-5 accuracy = 83.4%) for low frequencies of weight transfer (every 20, 10, or 5 epochs), which was lower than the average testing top-1 accuracy (48.3%, top-5 accuracy = 84.1%) for high frequency of weight transfer (every 4, 2, or 1 epoch, p < .01) (Table 3-4, Fig. 3-10).

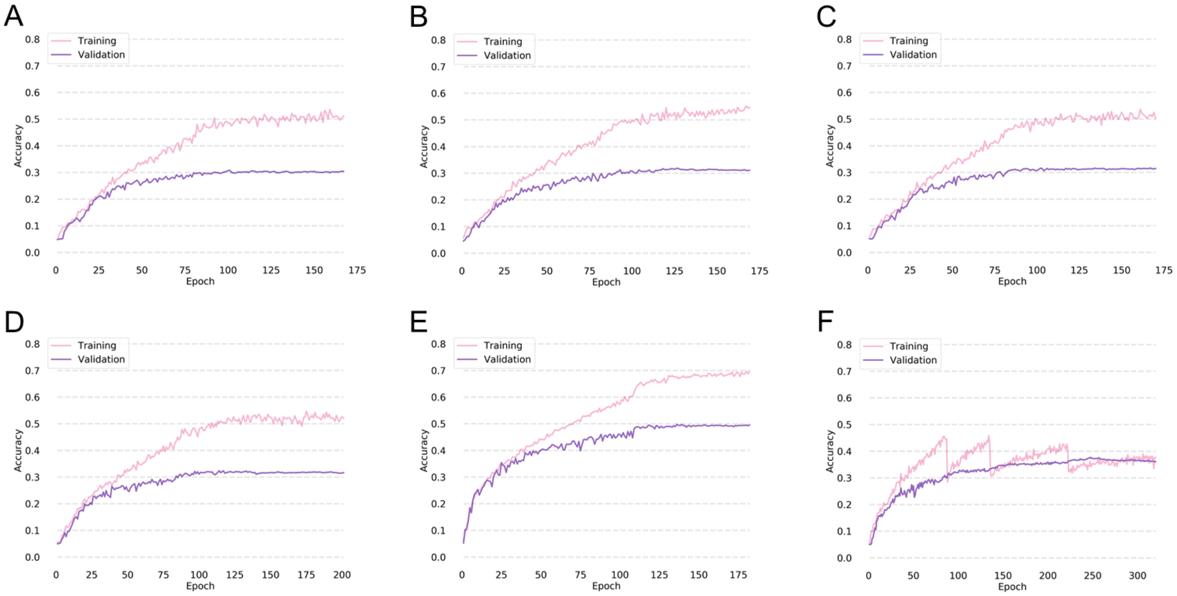


Figure 3-9. Training and validation accuracies during training on ImageNet dataset when trained on (A) Institution 1, (B) Institution 2, (C) Institution 3, and (D) Institution 4, (E) Centrally Hosted Training Heuristic, and (F) Single Weight Transfer Training Heuristic.

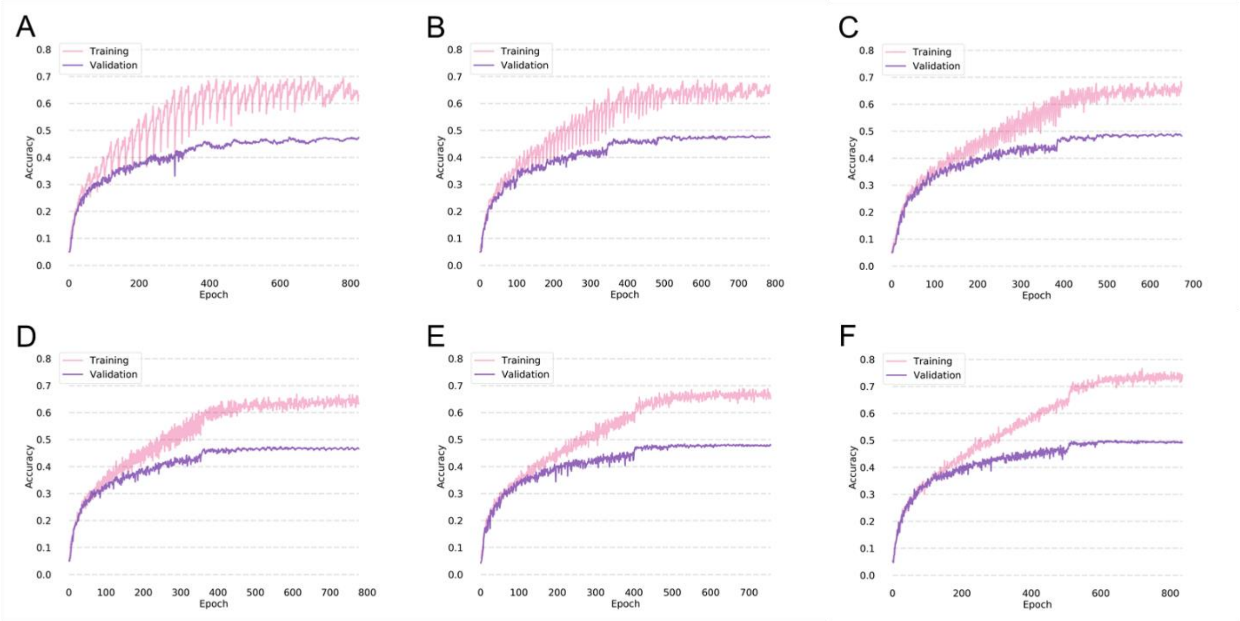


Figure 3-10. Training and validation accuracies during training on the ImageNet dataset with cyclical weight transfer with weight transfer frequencies of every (A) 20 epochs, (B) 10 epochs, (C) 5 epochs, (D) 4 epochs, (E) 2 epochs, or (F) every epoch.

3.4 Discussion

All training heuristics, either data sharing or model distribution, outperformed models trained only on one institution in terms of testing accuracy. This shows the benefits of collaboration among multiple institutions in the context of deep learning. Unsurprisingly, a model trained on centrally hosted data had the highest testing accuracy, serving as a benchmark for the performance of our various model sharing heuristics. In this study, we investigate if a model sharing heuristic can replace having the data be centrally hosted.

To overcome limitations in data-sharing, we tried several approaches – ensembling of single institution models, single weight transfer, and cyclical weight transfer. Ensembling of neural networks trained to perform the same task is a common approach to significantly improve the generalization performance. [147] In comparison, the concept of single weight transfer is very similar to that of transfer learning, which is derived from that idea that a model can solve new problems faster by using knowledge learned from solving previous problems in other domains. [148,149] In practice, this involves training a model on one institution's dataset and fine-tuning the model on a different dataset. If we consider each institution as a separate dataset, the model is trained on institution 1 and fine-tuned on institutions 2, 3, and 4. Both ensembling single institution models and single weight transfer resulted in higher testing accuracies than any single institution model for Kaggle Diabetic Retinopathy, DDSM, and ImageNet datasets. Single weight transfer outperformed ensembling models for the Kaggle Diabetic Retinopathy dataset while ensembling models and single weight transfer had the same testing performance for the DDSM dataset. For the ImageNet dataset, ensembling models outperformed single weight transfer.

The highest testing accuracies amongst training heuristics was cyclical weight transfer. On average, the testing accuracy of models trained with cyclical weight transfer was 1.9%, 2.5%, and 1.2% less than that of a model trained on centrally hosted data for the Kaggle Diabetic Retinopathy, DDSM, and ImageNet datasets, respectively. This means non-parallel distributed training produced model performance comparable to centrally hosted model performance, and parallel distributed training was not required to achieve this performance.

Furthermore, we find that a higher frequency of weight transfer had a higher testing accuracy than a lower frequency of weight transfer. For the Kaggle Diabetic Retinopathy dataset, the higher frequency of weight transfer had, on average, a 1.4% increase in testing accuracy compared to lower frequency of weight transfer. Similarly, for the DDSM dataset, a higher frequency of weight transfer had, on average, a 2.0% increase in testing accuracy compared to lower frequency of weight transfer. Finally, for the ImageNet dataset, a higher frequency of weight transfer had, on average, a 1.1% increase in testing accuracy compared to lower frequency of weight transfer. The disadvantage of having a higher frequency of weight transfer, however, is that it may be more logistically challenging and may add to the total model training time. In these cases, a lower frequency of weight transfer would still produce results that are comparable to that of a model trained on centrally hosted data. Lastly, we show that cyclical weight transfer is robust even when there was an institution with variability (either low-resolution images or few patients with class-imbalance), simulating a real-world scenario. We show that cyclical weight transfer performs similarly when the variable institution was introduced compared to when the variable institution is skipped entirely in terms of testing accuracy. In other words, variability did not significantly compromise the performance of the model with the cyclical weight transfer training heuristic.

In our experiments with 4 institutions, we show that we are able to achieve high model performance without having the data centrally hosted. We next investigated whether high model performance can be achieved when the performance of any single institution is no better than random classification. We divided 6000 patient samples from the Diabetic Retinopathy dataset into 20 institutions, each with 300 patient samples. Indeed, when we trained a model using data from one institution, the performance was no better than random classification. As we increased the number of collaborating institutions (using cyclical weight transfer), we observed an increase in testing accuracy. With all 20 institutions, cyclical weight transfer achieved a testing accuracy on par with centrally hosted data with all 6000 patient samples. This simulates a scenario where patient data are dispersed sparsely across many different institutions, and it is impossible to build a predictive model with data from any single institution. There are many situations (especially with rarer patient conditions) where no single institution has much patient data. In such cases, distributed learning can effectively utilize data from many institutions as long as the institutions are willing to distribute the model. In other words, if all institutions participate, they can, in essence, build a model capable of performing as if they had open access to all the data.

3.5 Limitations

One limitation is that our "institutions" were sampled from a single dataset (such as Kaggle Diabetic Retinopathy dataset) and thus, do not display much variability from one institution to the next. To address the possibility of variability, we performed experiments in which we altered one institution to either have low-resolution images or low numbers of patients with class imbalance. Further, in a follow-up study we assessed optimization strategies to account for the case when there is label distribution or quantity skew across multiple institutions.[7] Future studies can explore other types of heterogeneity, such as differences in data acquisition. Furthermore, for the Diabetic Retinopathy and DDSM datasets, the neural networks were trained to perform a binary classification problem. In practice, multi-label problems are commonplace but our work does not address how the added complexity would impact the various training heuristics. Future work can investigate the performance of distributed training heuristics in scenarios with multiple labels and more narrow decision boundaries. Also, we only investigated distributed learning in the context of a convolutional neural network. Distribution of models across institutions for other forms of deep learning, such as autoencoders, generative adversarial networks, and recurrent neural networks, warrant further study.[150–152] Lastly, future work will be on developing an open-source platform for distributed training. One key feature that is needed within this platform for cyclical weight transfer is that training at a given institute only begin after the training at the previous institute is completed.

3.6 Future directions

3.6.1 Other variants of distributed learning

Federated learning is a variant of distributed learning in which training among institutions is orchestrated by a central server.[133] Under this paradigm, models are trained locally at each institution and either gradient updates (federated stochastic gradient descent) or model weights (federated averaging) are sent to the central server.[153,154] The model in the central server is then updated and the weights are then sent back to the institutions to update their local copy of the model. A key challenge with federated stochastic gradient descent is waiting for synchronous updates from each institution. Specifically, the model weights in the central server can only be updated and sent back to the institutions after gradient updates are received from all institutions.[155] Given that each institution is likely to have different compute and communication infrastructure, this process is rate-limited by the slowest institution. A workaround is to perform asynchronous stochastic gradient descent in which each institution asynchronously grabs the most up-to-date model weights from the central server, computes gradients of the loss, and then send the gradients back to the central server.[156] The downside of asynchrony is while a specific institution is calculating the gradient, the model weights in the central server may be updated by other institution, resulting in the gradients from the specific institution being calculated with outdated model weights. These gradients are thus termed stale resulting in convergence with worse performance.[155] A compromise between fully synchronous or fully asynchronous approaches is to have partial synchrony, in which the central server waits for update from institutions until a certain point, after which the updates from straggler institutions are discarded.[155] An alternative approach is federated averaging, in which local model weights are sent to the central server after a specified number of training iterations and averaged. The averaged model weights are then sent back to the local institutions to update the local copy of the model.[154] The advantage of this approach is that communication with central server is only performed after a specified number of training iterations (which consists of many gradient updates), which can more communication efficient, depending on the frequency of averaging. However, federated averaging faces the same synchrony challenge as federated stochastic gradient descent in terms of being rate limited by the slowest institution. Recently federated averaging has been shown to be capable of achieving near centrally hosted performance under a simulated setting, for the task of brain tumor segmentation.[131]

Split learning is based on the idea that the layers of a neural network can be divided piecewise at specific layers (termed cut layers) between institutions and the central server.[133,157,158] Raw patient data is never shared, but rather the outputs of cut layers (termed smash data) during forward propagation and the gradients of cut layers during backpropagation.[133] Although there are many possible configurations, the most relevant one for medical applications is U-shaped (boomerang) split learning, designed for a scenario in which both input data and labels cannot leave the institution.[158] In this paradigm, each institution has their own beginning and end layers of a neural network. The intermediate layers are shared between all institutions. At each iteration of training, the image is fed through the local beginning layers, then fed through the intermediate layers on the central server, and lastly through the local end layers. The loss and gradient are then calculated and the weights are updated through backpropagation, this time moving through all the layers in reverse. The main advantage of split learning is the ability to defer a portion of the computation to the central server.[158,159] This decreases the computational resources needed at each institution, which would benefit institutions with limited resources, such as smaller community hospitals. Also, depending on the dataset size contributed per client, number of clients and the model size, the communication requirements of split learning can be more favorable than those of federated averaging.[158,160] A recent study has shown the potential of split learning for achieving centrally hosted performance for healthcare applications.[161]

Compared to cyclical weight transfer, federated learning and split learning have differing synchrony, communication, and hardware requirements. Furthermore, time to convergence and performance at convergence may differ. Each method provides different advantages and disadvantages that warrant further head-to-head comparison in real-world healthcare use cases.

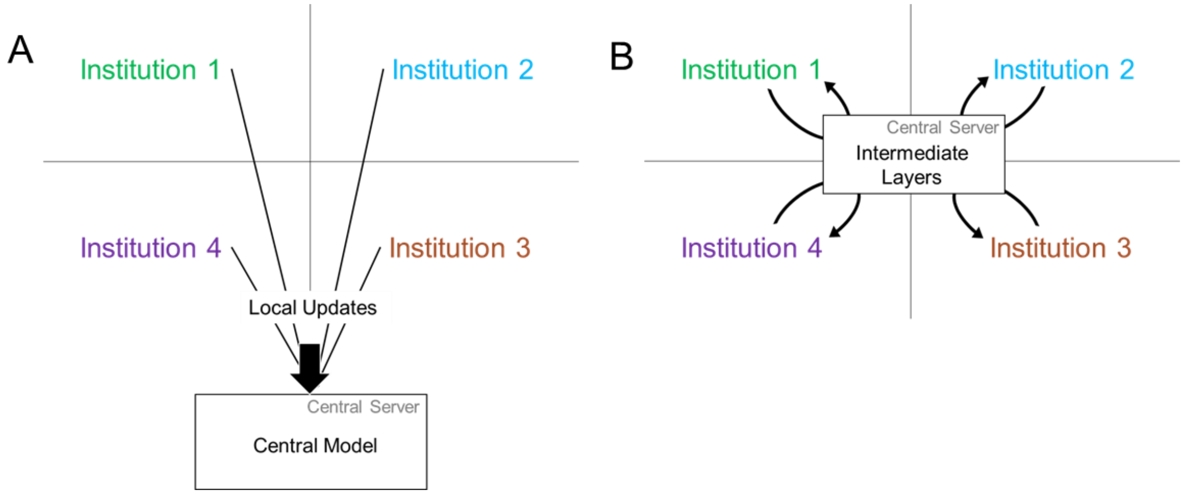


Figure 3-11. Other variants of distributed learning include (A) federated learning, in which local model weights or gradients are average on a central server and (B) split learning, in which intermediate layers are shared across institutions and smashed data is transferred during forward propagation and gradients are transferred during back propagation.

In model averaging, separate models are trained for each split of the data and the weights of the model are averaged every few mini-batches.[162] In asynchronous stochastic gradient descent, separate models are trained for each split of the data and the gradients of each separate model are transferred to a central model.[156] However, these methods were developed with the intention of optimizing training speed. Although applying such data parallel training methods in a multi-institution study in which data is not exchanged between institutions is possible, they also represent a significant logistical challenge. Specifically, training would have to take place in parallel across all institutions. This would be especially challenging if institutions have drastically different network connection speeds or deep learning hardware. While non-parallel methods of distributed training may be slower than parallel methods, they would avoid the logistical challenges.

3.6.2. Handling data heterogeneity

In this chapter, we showed that cyclical weight transfer is robust to low-resolution images or low numbers of patients with class imbalance at a single institution. In a follow-up study, we detail optimization approaches for label distribution or quantity skew across multiple institutions.[7] However, in a real-world use case, there will likely be a combination of multiple types of heterogeneity (quantity skew, feature distribution skew, label distribution skew, and concept shift) across all institutions. As such, further approaches need to be considered to achieve high performance. One hurdle is the high prevalence of Batch Normalization (BatchNorm) layers in modern neural network architectures.[137,163,164] BatchNorm layers stabilize neural networks by channel-wise normalization of intermediate inputs with the mean and standard deviation of each mini-batch, which mitigates divergent effects of large gradient updates and smooths the optimization landscape.[165,166] At inference time, an estimate of the global mean and standard deviation is used. This can be problematic in a non-IID setting with federated learning because the training and validation distributions differ, resulting in differing normalization during training and validation and thus, lower validation performance.[132] Hsieh et al provides evidence that much of the loss in performance due to BatchNorm can be partially recovered by replacing the BatchNorm layers with Group Normalization (GroupNorm) layers.[132] GroupNorm layers normalize by group, which is defined as a prespecified number of adjacent channels for each individual input (as opposed to on a mini-batch basis).[167] Another hurdle is the prevalent use of momentum in neural network optimizers, which improves convergence of networks.[1,168] However, it is unclear how to incorporate momentum into distributed learning.[133] Work by Yu et al demonstrates that letting each institution have its own momentum buffer followed by periodic global averaging of the buffers improves accuracy of the final model compared to resetting the buffers to 0 during each round of federated averaging.[169]

One critical barrier to optimizing of distributed training is dealing with catastrophic forgetting, a phenomenon in which sequential training of a model on a new task results in "forgetting" of previously learned knowledge for previous tasks.[170] Although the task is the same across all institutions in most healthcare applications, if the dataset is non-IID across institutions, it is possible that learning the model may "forget" what it learned at other institutions when training at a given institution. This is of particular concern in cyclical weight transfer, in which training occurs at each institution in sequence.[5] Indeed, overall model performance is decreased with decreasing frequency of weight transfer.[5] Catastrophic forgetting is also a concern in the context of synchronous distributed learning (such as federated averaging) if the gradient updates at one institution are anti-parallel to those from another institution. Some proposed approaches to dealing with catastrophic forgetting may be to slow down learning on weights that are important for other institutions when updating model weights for a given institution, masking trainable weights differentially at each institution, or to update the model weights orthogonally for each institution.[171–173]

Another approach to dealing with data heterogeneity is to train "personalized" models for each institution as opposed to a single global model for all institutions, also known as domain adaptation.[133] These "personalized" models can be adapted from a global model that performs reasonably well among all institutions. For example, one strategy could be to have common weights for convolutional layers of the neural network but have institution-specific BatchNorm layers.[174] Another strategy would be to perform unsupervised domain adaptation of the global model to a specific institution through adversarial training.[175] Zhao et al gives upper and lower bounds on conditions required to reduce model error on target domain when adapting from a source domain.[176] The bounds are based on Jenson-Shannon divergences between labels distributions in source and target domain as well as between intermediate representations learn by the deep learning network over the source and target domain of data. Once these "personalized" models are trained, there will be several models that can be used for inference on new institutions. Model selection strategies, based on similarity of the data distribution of the new institution with the data distribution used to train each of the “personalized models”, can be used to select the optimal model.[177,178]

Alternatively, datasets at institutions can be augmented to make the data distribution more IID-like. For example, in the presence of imbalance in the distribution of class or patient characteristics, data from the minority class or characteristic can be augmented via synthetic oversampling.[179] If one institution has less data than another institution, the institution with less data can augment its data with geometric transforms, mix-up, or generative adversarial networks (GANs), assuming that such augmentation doesn’t cause a shift in the institution’s overall distribution.[180–183] If scanner types or image sequences differ across institutions, data at each institution can be augmented for acquisition diversity using supervised approaches.[184–186] In summary, these heterogeneity present a critical hurdle to the deployment of distributed deep learning methods. Importantly, these challenges not unique to distributed machine learning, but relevant for multi-institution machine learning as a whole.

3.6.2 Patient privacy

One of the key motivations of distributed deep learning is the protection of patient privacy. Most institutions require researchers to deidentify patient data before they are used for training. This removes obvious patient identifiers such as name, medical record numbers, date of birth, and date of hospital visit. It is important to note that patient information is still embedded in the clinical variables, lab tests, and medical imaging. Distributed learning provides a method of training deep learning models without sharing raw patient data. However, this is not equivalent to full protection of patient privacy as there is still component data being shared: model weights (model ensembling, cyclical weight transfer, federated averaging), gradients (federated stochastic gradient descent), and smashed data (split learning). A tech-savvy attacker might infer sensitive information about the training data or, in the worst-case scenario, reconstruct the training data itself from the shared component data.[187–189] As such, additional protections need to be put into place to ensure privacy among participating institutions.

One framework for such protection is differential privacy (DP). At the core of DP is the concept of a privacy budget, which is the maximum increase to the risk of an individual's privacy.[190] Alternatively, the privacy budget is how much of an individual's privacy that the neural network can use for training. In practice, DP optimization involves clipping the gradient followed by addition of gaussian noise at each training step. Training is discontinued once the privacy budget is exhausted, regardless of whether the desired level of performance is reached.[191] Utilization of DP comes with an important tradeoff – there is an inversely proportional relationship between the privacy budget and model performance.[159] That is, the more stringent the protection, the lower the model performance. Recently, DP has been applied to train models for healthcare applications.[192,193] DP can also be used to train GANs that can subsequently be shared for model training.[194]

Homomorphic encryption is an approach that allows performing of mathematical operations directly on the encrypted data (ciphertext).[133] The allowed operations include addition and multiplication, but other operations (such as activation functions) can be approximated using higher degree polynomials, Chebyshev polynomials, and Taylor series.[159] Patient data or smashed data can be homomorphically encrypted before it is sent to the central server for model training or inference.[133,195,196] Alternatively, model weights can be homomorphically encrypted before it is sent to the central server for aggregation.[133] The major drawback of homomorphic encryption is the need for specialized hardware and extensive computational resources, limiting the scalability of the method.[159,195] As such, much of the work on homomorphic encryption has been focused on shallow architectures that do not represent the deep architectures used for modern healthcare applications.[159,195,196]

One concern that is specific to split learning is the correlation between raw data and smashed data, resulting in leakage of information. To reduce this leakage, a variant of split learning, called NoPeek, utilizes a decorrelation approach based on distance correlation between raw and smashed data as part of the loss function during optimization.[133,189] This approach has been shown to protect again information leakage while maintaining high model performance.[189] This is especially useful when the cut layer is very early in the neural network when the correlation between raw and smashed data is high.[189] The exact tradeoff between the use of NoPeek and model performance in a variety of medical deep learning scenarios is still under investigation.

3.7 Conclusions

In this chapter, we address the question of how to train a deep learning model without sharing patient data. We found that cyclical weight transfer performed comparably to centrally hosted data, suggesting that sharing patient data may not always be necessary to build these models. This finding has applications for any collaborative deep learning study. There are other methods for distributed learning as well that became popularized in parallel with our work, namely federated learning and split learning. Each method provides different advantages and disadvantages that warrant further study in real-world healthcare use cases. A large hurdle to such validation is dealing with the presence of data heterogeneity within and across institutions, which present challenges in optimization, generalizability, and catastrophic forgetting. This hurdle is compounded by the need for further protection of patient privacy, which induce tradeoffs in performance and computational complexity. Studies on the synergy between methods of distributing training, handling data heterogeneity, and protecting patient privacy provide an avenue for impactful future work.

4 Model design and the impact on performance

4.1 Introduction

Breast cancer is a leading cause of death among women in the US, with the expected number of deaths to be over 41,000 in 2019.[197] Early mammographic screening has resulted in a decrease in breast cancer mortality.[198,199] The correct mammographic interpretation of breast density, which measures extent of fibroglandular tissue, is important in the assessment of breast cancer risk as there is increased risk with increased density.[200,201] Furthermore, the identification of dense breast may stratify patients who may have masked cancers and may benefit from additional ultrasound and/or MR imaging. As such, there is now legislation in many states that patients must be notified of their breast density after mammography.[202]

Qualitative assessment by means of the widely used Breast Imaging Reporting and Data System (BI-RADS) include four categories: a) almost entirely fatty, b) scattered fibroglandular densities, c) heterogeneously dense, or d) extremely dense.[203] These criteria are subjective, resulting in inter-rater variability among radiologists. A study by Sprague et al. showed that the likelihood of any given mammogram being rated as dense (heterogeneously dense and extremely dense) is highly dependent on the interpreting radiologist, with the percentage ranging from 6.3%-84.5%.[204] Similarly, commercially available software shows a wide range of agreement with clinical experts and the probability of dense classification is dependent on the specific software used.[205,206] This inter-rater variability, and even inter-software variability, may confer undue patient anxiety and potential harm to the patient, i.e. possible unnecessary supplemental screening examinations.

As such, there has been interest in using automated approaches to improve accuracy and consistency of breast density assessment. Commercial software utilize quantitative imaging features to assess breast density, with mixed agreement with radiologist interpretation.[206] Deep learning methods have yielded state-of-the-art results in a wide range of computer vision tasks without the need for domain-inspired hand-crafted imaging features. Moreover, recent studies have shown the potential of deep learning in medical fields such as dermatology, ophthalmology, and radiology.[40,42,207] A recent study from Lehman et al. demonstrates the utility of deep learning for mammographic density assessment in clinical practice at a single institution/mammography system. [202] Here, we further this work by validating the deep learning approach on a multi-institutional imaging cohort with a variety of digital-mammography systems. Furthermore, we provide an in-depth analysis of how choice of data, model, and training parameters affects algorithm performance. In addition to that, we investigate the generalizability of models across different digital-mammography data formats. Lastly, we deploy our system at the American College of Radiology (ACR) 2019 Annual Meeting for a crowdsourced evaluation.

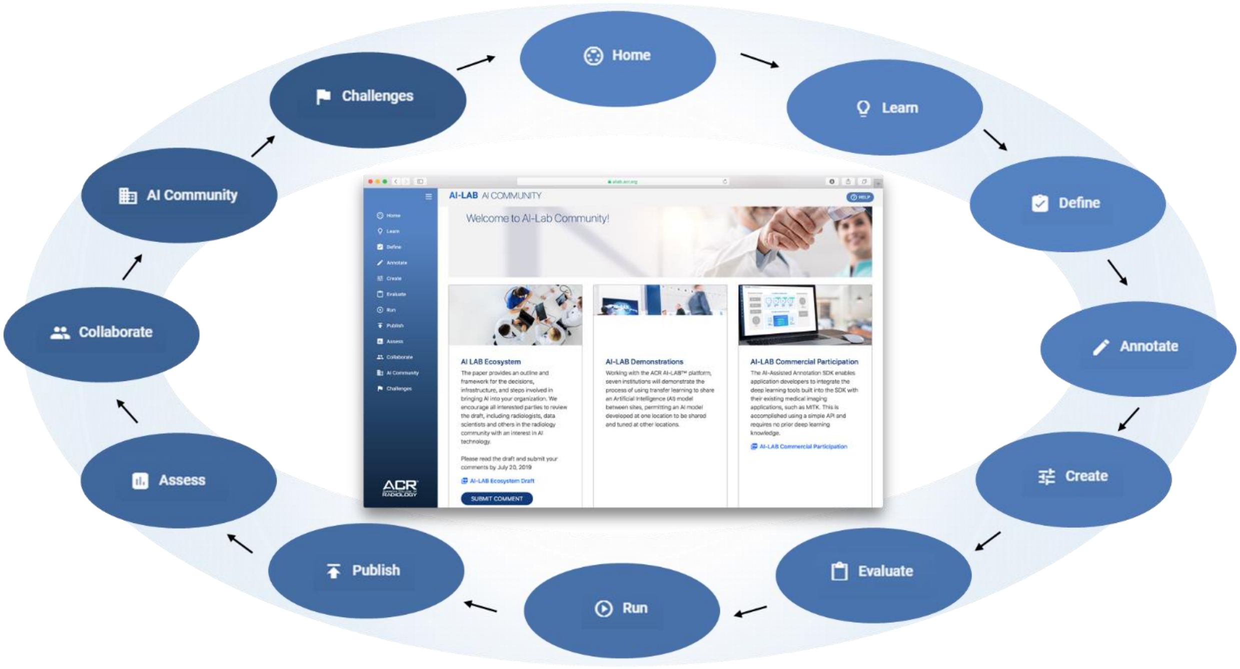
We highlight several fundamental features needed for artificial intelligence democratization: First, we demonstrate the possible data, model, and training parameters that can influence the performance of the model. We also show importance of diverse training data for model generalizability, supporting collaborative development of algorithms across institutions. Lastly, we show how a crowdsourced annotations can be used to evaluate algorithm performance.

4.2 ACR AI-LAB

Despite significant research into the applications of AI, there is currently limited use of AI in clinical care. Key to the success of these algorithms are three components: 1) the availability of large quantities of diverse, well-annotated patient data, 2) clinical professionals who can drive direction, validation, and translation, and 3) data scientists who can design, train, and deploy such algorithms. Unfortunately, such synergy is only accessible within certain academic institutions. Of note, the availability of publicly-available, high-quality data of sufficient size has been known to be a bottleneck for progress in AI.[208] To date, radiology professionals, who have the requisite domain expertise to make AI relevant for clinical use, have not been able to widely participate in AI development because of limited access to AI computational solutions and the complexity of AI computational architecture. On the other hand, data scientists who are not working closely with radiology professionals are building algorithms that are accurate yet clinically irrelevant or not useful. The widespread availability of these three components to all individuals and institutions would catalyze AI research and expedite model integration into the clinical workflow.[209]

Even once the model is trained, there are many foregoing challenges. First, end-users for these models, such as radiologists and pathologists, may not fully understand how these models were trained. Critically, this may lead to a general misunderstanding for why and when a model will fail. Furthermore, models need to be rigorously validated to ensure that they are not biased to the technical imaging specifications and patient population of the training data. In fact, only 6% of AI studies report external validation.[80] Consequently, models trained on data from only one institution may not achieve high performance (due to limited size or diversity of the training dataset), and moreover may not generalize well to data from other institutions. Thus, there exists a need to enable multi-institutional, collaborative approaches for the purpose of creating robust models, while simultaneously ensuring patient privacy.[5]

To this end, the ACR presents AI-LAB, a framework for democratization of AI which was developed in tandem with study presented in this chapter. The AI-LAB provides an online interface for radiologists, radiation oncologists, medical physicists, and data scientists to work together, both within institutions and across institutions. Among its core functions, the AI-LAB provides education on AI, tools for data curation, annotation, model development, and collaboration, all in a code-free environment. Recently, the AI-LAB was used in a resident challenge, where radiology residents experimented with the built-in options for model design and evaluated the effect of performance of a deep learning mammographic breast density model. Such a platform allows the greater radiological community to overcome many challenges faced within AI and lays the foundation towards the shared goal of improving patient care.



Welcome to AI-Lab Community!

AI LAB Ecosystem

The paper provides an outline and framework for the decisions, infrastructure, and steps involved in bringing AI into your organization. We encourage all interested parties to review the draft, including radiologists, data scientists and others in the radiology community with an interest in AI technology.

Please read the draft and submit your comments by July 20, 2019

AI-LAB Ecosystem Draft

AI-LAB Demonstrations

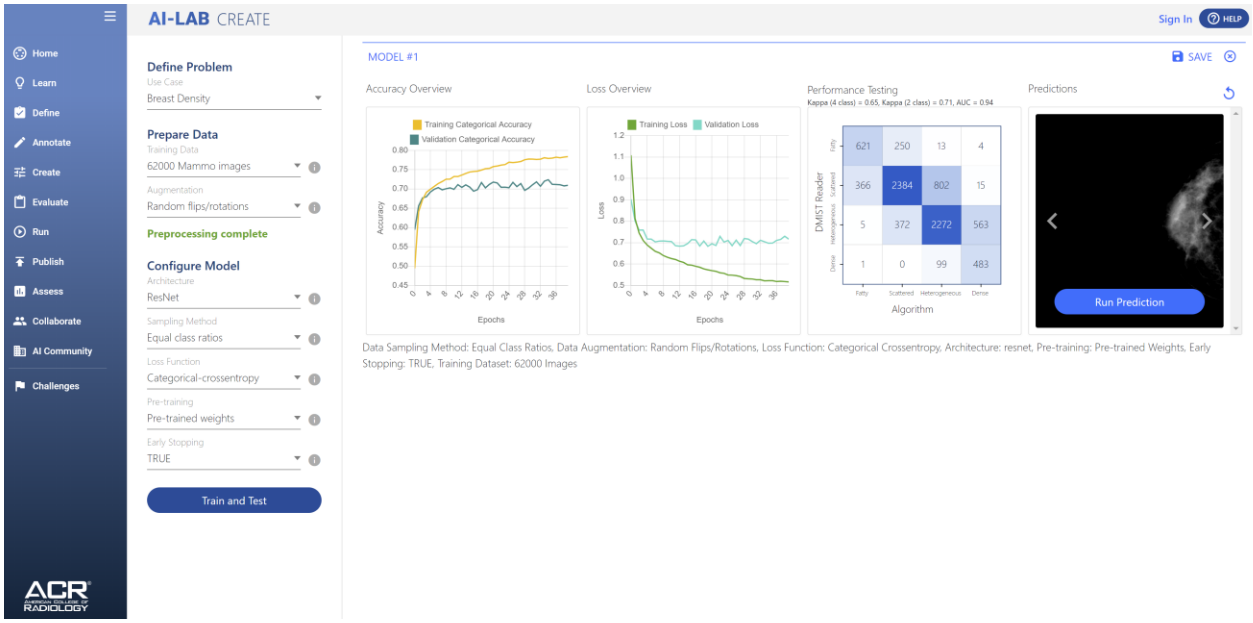
Working with the ACR AI-LAB™ platform, seven institutions will demonstrate the process of using transfer learning to share an Artificial Intelligence (AI) model between sites, permitting an AI model developed at one location to be shared and tuned at other locations.

AI-LAB Commercial Participation

The AI-Assisted Annotation SDK enables application developers to integrate the deep learning tools built into the SDK with their existing medical imaging applications, such as MITK. This is accomplished using a simple API and requires no prior deep learning knowledge.

AI-LAB Commercial Participation

Figure 4-1. A schematic of the AI-LAB Ecosystem, which consists of 11 modules for education, annotation, model development, model evaluation, data and model sharing, and distributed training. Source: ACR



Define Problem

Use Case: Breast Density

Prepare Data

Training Data: 62000 Mammo images

Augmentation: Random flips/rotations

Preprocessing complete

Configure Model

Architecture: ResNet

Sampling Method: Equal class ratios

Loss Function: Categorical-crossentropy

Pre-training: Pre-trained weights

Early Stopping: TRUE

Kappa (4 class) = 0.65, Kappa (2 class) = 0.71, AUC = 0.94

	Fatty	Scattered	Heterogeneous	Dense
Fatty	621	250	13	4
Scattered	366	2384	802	15
Heterogeneous	5	372	2272	563
Dense	1	0	99	483

Data Sampling Method: Equal Class Ratios, Data Augmentation: Random Flips/Rotations, Loss Function: Categorical Crossentropy, Architecture: resnet, Pre-training: Pre-trained Weights, Early Stopping: TRUE, Training Dataset: 62000 Images

Figure 4-2. Example of using AI-LAB to train and evaluate a deep learning mammographic breast density model. Source: ACR

4.3 Materials and methods

4.3.1 Patient cohort

Digital screening mammograms were retrospectively obtained through the Digital Mammographic Imaging Screening Trial (DMIST), the details of which were previously published.[210] In summary, women were recruited to the study across 33 sites and underwent digital mammography from various digital-mammography systems. The protocol was approved by the institutional review boards at all sites and all patients gave written informed consent. Each site had a lead radiologist that trained the sites' other radiologist readers.

Each examination was interpreted by a single radiologist from a cohort of radiologists using ACR BI-RADS breast density lexicon (Category a: fatty, Category b: scattered, Category c: heterogeneously dense, Category d: extremely dense).[203] All images were previously de-identified before this study. The mammograms were saved in DICOM format with 4 different image data formats, corresponding to different digital-mammography systems or different versions of the same system (Table 4-1): 12 bit Monochrome 1 (30.3%), 12 bit Monochrome 2 (11.2%), 14 bit Monochrome 1 (58.0%), and 14-bit Monochrome 2 (0.5%). 14-bit Monochrome 2 images were excluded to ensure that each image data format included in our study had adequate representation for training of our deep learning model. Monochrome 1 images indicates images were reversed from conventional intensity representation in which higher intensities indicate higher opacity. Monochrome 2 images were in conventional intensity representation. Monochrome 1 images were inverted as part of preprocessing.

Our final patient cohort consisted of 108,230 digital screening images from 21,759 patients. The demographics of the patient cohort is shown in Table 4-2. We divided this cohort on the patient level into training (n = 62,316 images from 12,158 patients), validation (n = 6,978 images from 1,351 patients), and testing sets (n = 38,936 images from 8,250 patients). The training set was used to develop the model and the validation set was used to assess model performance during training to prevent overfitting. The test set was unseen until the model training was complete.

12 Bit Monochrome 1	12 Bit Monochrome 2	14 Bit Monochrome 1
Senoscan (99.9%) Kodak Lumiscan 75 (.1%)	Senograph (93.8%) Other (6.1%) Mammo-Clinical (.1%)	Senograph (94.1%) Mammo-Clinical (5.9%)

Table 4-1. Breakdown of data format by digital mammograph system.

	Training (n = 62316)	Validation (n = 6978)	Testing (n = 38936)
Age (median years, IQR)	46 (53-61)	46 (53-61)	47 (53-61)

Female (%)	100	100	100
Race White	50414	5622	30845
Black or African American	8389	925	5733
Hispanic or Latino	2273	289	1416
Asian	819	62	633
American Indian or Alaska	63	8	19
Other or Unknown	358	72	290
Radiologist-assessed breast density			
Fatty	6980 (11.2%)	873 (12.5%)	4575 (11.8%)
Scattered	27733 (44.5%)	2985 (42.8%)	17191 (44.2%)
Heterogeneously dense	23987 (38.5%)	2753 (39.5%)	14585 (37.5%)
Extremely dense	3616 (5.8%)	367 (5.3%)	2585 (6.6%)

Table 4-2. Summary of demographics in the patient cohort with regard to age, sex, race, and breast density.

4.3.2 Image preprocessing

For Monochrome 1 images, image intensity values were inverted to ensure all images transitioned from radiolucent to radiopaque as intensity value increased. Because images of each image format contained different intensity distributions (Fig. 4-3), the intensity of each image was normalized by the mean and standard deviation of the population mean and standard deviation of that image format. [202] To ensure proper input size to the pre-trained neural network architectures that were used, the images were downscaled with linear interpolation and replicated to create three channels. The final preprocessed image size was 224x224x3 as to be able to use ImageNet pretraining.

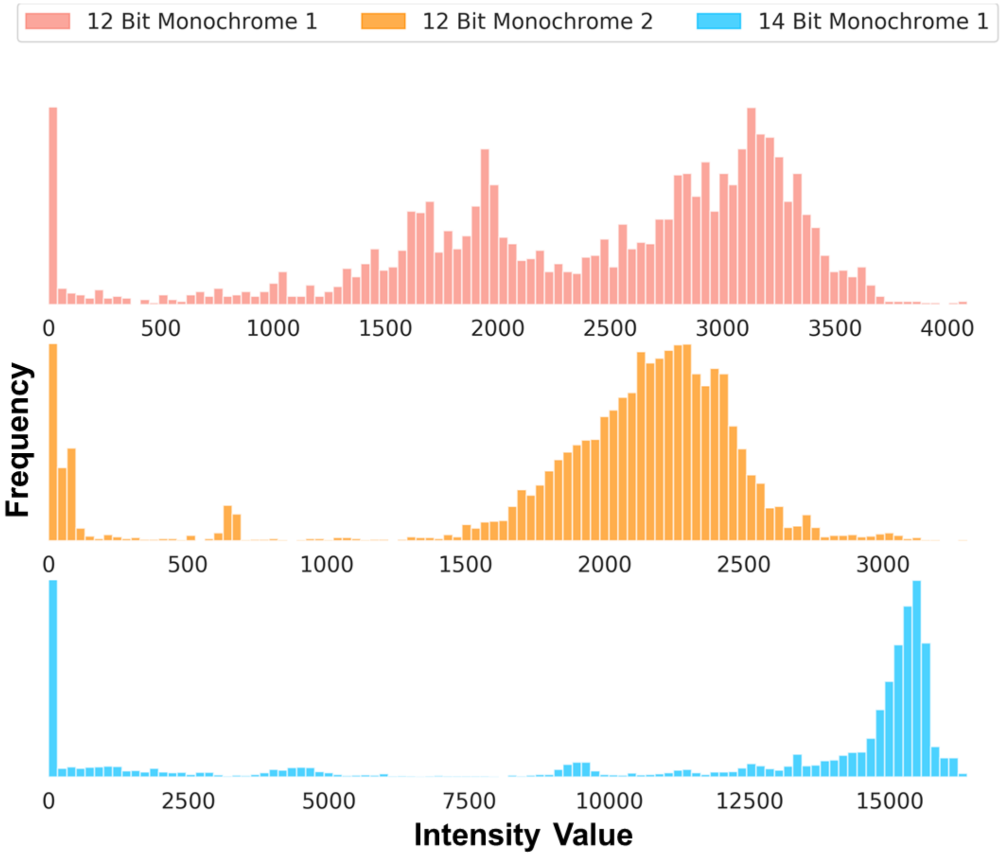


Figure 4-3. Visualization of an intermediate layer of the trained neural network for 3000 images in the testing set, color-coded by (A) image format and (B) radiologist interpretation of breast density.

4.3.3 Training

Neural network models were implemented in DeepNeuro with Keras/TensorFlow backend. [10,138,211] Models were trained using images from a single view. All models were optimized using the Adam optimizer with an initial learning rate of $1*10^{-6}$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$.[145] All model parameters were made trainable, regardless of whether the model was pretrained or randomly initialized. The learning rate was decreased by a factor of 10 when the loss on the validation set did not improve for 10 consecutive epochs. The network was trained until the loss on the validation set did not improve for 20 consecutive epochs. The model with the lowest validation set loss during training was saved as the final model. At prediction time, the probability of each density class was predicted using the neural network for each image individually. To combine predictions from all images (across all views) from a given patient into a patient-level assessment, the probabilities for all images were averaged. The averaged probabilities were then used to determine the predicted breast density class.

4.3.4 Experiments on data, model, and training parameters

We investigated the effect of data, model, and training parameters on algorithm performance. 13 different models were trained with varying parameters. A schematic of the various experiments investigating data, model, and training parameters are summarized in Fig. 4-4. To investigate the effect of training set size, we utilized various different training set sizes and assessed the resulting performance on the test set. We tested four commonly used neural network architectures, each of which differ in number of layers and design: ResNet50 (23,542,788 parameters), DenseNet121 (6,957,956 parameters), InceptionV3 (21,776,548 parameters), and VGG16 (14,716,740 parameters).[137,212–214] We also investigated the benefit of pretraining by comparing ImageNet (a large computer vision dataset of natural images) pretrained versus random initialization.[215] A variety of cost-functions were also utilized (categorical cross-entropy, mean absolute error, mean squared error, and ordinal regression) in order to assess the effect of objective function (and their underlying assumptions of the nature of the labels) on performance.[216] The last layer of the neural network was modified to accommodate the dataset and cost function: four-unit dense layer with softmax activation (categorical cross-entropy), single-unit dense layer with linear activation (mean absolute error and mean squared error), and four unit-dense layer with sigmoid activation (ordinal regression). The training set was augmented in real time by means of random horizontal/vertical flips (50% probability of each) and random rotations (0-45°). At each mini-batch, images from each breast density class were sampled with either random (weighting in the empirical density class distribution) or equal class (weighting each density class equally) probability to assess the effect of class weighting on performance. We also evaluated the effect of model ensembling by averaging the output of 2-4 trained models of the same architecture (ResNet50). Model ensembling describes the process by which several independently trained models are combined to improve performance.[141] The default model utilized 100% of the training set, ImageNet pretrained weights, ResNet50 architecture, no ensembling, categorical cross-entropy loss function, augmentation, and equal class sampling. Only one parameter was modified at a time in the experiments, keeping all other parameters the same as the default model (*ceteris paribus*).

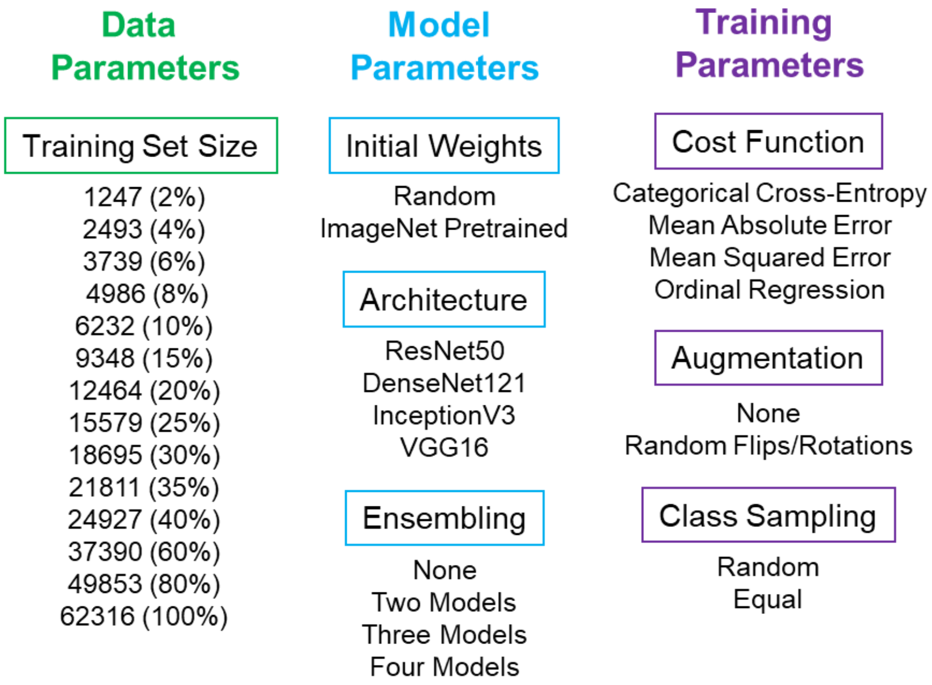


Figure 4-4. A summary of all the data, model, and training parameter experiments performed.

4.3.5 Experiments on image data formats

To visualize the differences in intensity distributions across image formats, histograms of preprocessed images from the testing set were generated. The dimensionality of histograms ($x \in \mathbb{R}^{100}$) were then reduced to a 2-dimensional projection and plotted to inspect for similarity across image formats.[217] The effect of image format of training images on generalizability of models was investigated. We trained ResNet50 models using 12 bit Monochrome 1 images only, 12 bit Monochrome 2 images only, 14 bit Monochrome 1 images only, and all images. The performance of these models for each image format was then assessed. Projections of an intermediate output from the penultimate layer of the neural network were also plotted for images in the testing set using a model trained on all images to evaluate the learned features learned by the deep learning model. The dimensionality ($x \in \mathbb{R}^{1000}$) was then reduced to 2 using Uniform Manifold Approximation and Projection (UMAP) and plotted to inspect for similarity across image formats.[217]

4.3.6 Crowdsourcing assessment

As further evaluation of our breast density algorithm, we deployed an annotation workstation at the ACR 2019 Annual Meeting. Attendees of all levels (researchers, medical students, residents, radiologists) were invited to perform annotations on a subset of images within our patient cohort. Representative images of all breast density classifications from the BI-RADS manual were provided to attendees during annotation. Attendees were able to inspect all images (all views available) from a given patient study and were asked to provide a BI-RADS breast density assessment. In total, 3,649 annotations were performed on 1083 patient studies by 17 raters (Demographics summarized in Table 4-3). On average, there were 3 annotations per patient study and each rater performed 215 annotations. Consensus of the crowd was determined by majority vote, with random tiebreak. In our analysis, we looked at agreement between crowd and radiologist annotation as well as crowd and algorithm (ResNet50),

	N
Experience	
Radiologist (Breast)	3
Radiologist (Other)	10
Resident	2
Student	2
Read Mammograms	
No	10
Yes	7

Table 4-3. Demographics of participants of the crowdsourcing assessment.

4.3.7 Statistical analysis

Agreement between raters was assessed via linear κ coefficient across the four breast density categories in the testing set (4-class κ). For reference, a κ of 0.21-.40, 0.41-0.60, and 0.61-0.80 represents fair, moderate, and substantial agreement, respectively.[218] Agreement between raters was also assessed via linear κ coefficient for non-dense (class a and b) vs. dense breast (class c and d) based on the categorization of notification requirements in most states. Experiments on data, model, and training parameters as well as image data format were repeated five times to calculate confidence intervals. For the crowdsourcing assessments, confidence intervals were calculated with non-parametric bootstrapping. Chi-squared test was used to compare the distributions of predicted labels between experiments. An unequal variances t-test at a significance level of p = .05 was used for statistical comparisons of model performance.

4.4 Results

4.4.1 Effect of data parameters on performance

The performance of training set size on testing set performance was investigated, showing that κ coefficient increases as the training set size increases. When 2% (n = 1247 images) of the training set was used, the mean 4-class κ was .563 (95% Confidence Interval, CI, .551-.576). In contrast, when using 100% (n = 62,316 images) of the training set, the mean 4-class κ was .660 (95% CI .657-.664) (Fig. 4-5). There was a statistically significant difference between the performance of using 2%-60% and 100% of the training set (t-test p < .05). There was no difference in the performance of using 80% and 100% of the training set (p = .291).

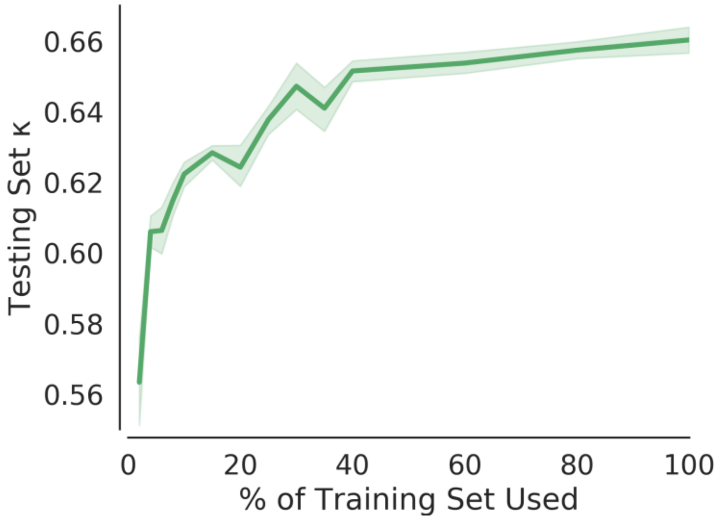


Figure 4-5. Performance on the testing set (measured by 4-class κ agreement with radiologist interpretation) increased as the percentage of training set used. The 95% confidence interval is plotted in light green.

When 2% (n = 1247 images) of the training set was used, the mean 2-class κ was .642 (95% Confidence Interval, CI, .629-.655). In contrast, when 100% (n = 62,316 images) of the training set was used, the mean 2-class κ was .718 (95% CI .712-.723). There was a statistically significant difference between the performance of using 2%-40% and 100% of the training set ($p < .05$). There was no difference in the performance of using 60% or 80% and 100% of the training set (p = .054 and .291, respectively).

4.4.2 Effect of model parameters on performance

The number of epochs until model convergence for randomly initialized weights (68.2, 95% CI 47.2-91.2) was greater than for ImageNet pretrained weights (38.2, 95% CI 35.0-41.6), although not statistically significant (p = .086).

The mean 4-class κ of models with randomly initialized weights was .327 (95% CI .273-.384), compared to ImageNet pretrained weights .660 (95% CI .657-.664, $p < .001$) when using the full training set (Fig. 4-6). In the experiments assessing model architecture, the mean 4-class κ of ResNet50, DenseNet121, InceptionV3, and VGG16 was .660 (95% CI .657-.664), .650 (95% CI .640-.659), .644 (95% CI .635-.652), and .660 (95% CI .658-.664), respectively. There was no statistically significant difference between the performance of the various architectures. The mean 4-class κ of no ensembling, ensembling two models, ensembling three models, and ensembling four models was .660 (95% CI .657-.664), .665 (95%

CI .664-.666), .666 (95% CI .666-.667), .667 (95% CI .666-.668), respectively. The performance of ensembling four models and three models was greater than that of no ensembling (p = .041 and .036, respectively).

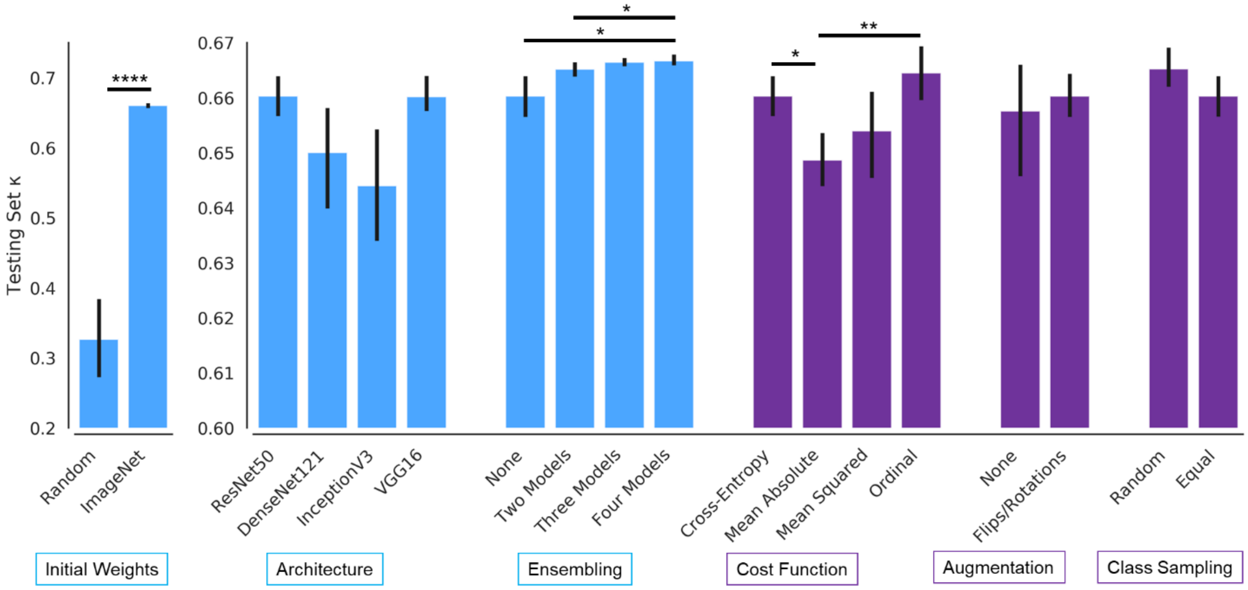


Figure 4-6. Effect of model and training parameters on testing set 4-class κ agreement with radiologist interpretation. Black lines denote 95% confidence interval. P-values are denoted by *p < .05, **p < .01, ****p < .001

The mean 2-class κ of models with randomly initialized weights was .453 (95% CI .389-.533), compared to ImageNet pretrained weights .718 (95% CI .712-.724, p < .005. In the experiments assessing model architecture, the mean 2-class κ of ResNet50, DenseNet121, InceptionV3, and VGG16 was .718 (95% CI .712-.724), .720 (95% CI .717-.723), .719 (95% CI .714-.724), and .722 (95% CI .719-.724), respectively. There was no statistically significant difference between the performance of the various architectures. The mean 2-class κ of no ensembling, ensembling two models, ensembling three models, and ensembling four models was .718 (95% CI .712-.724), .721 (95% CI .719-.724), .722 (95% CI .721-.724), .723 (95% CI .721-.724), respectively. There was no significant difference between the performance of ensembling models and that of no ensembling (p > .05).

4.4.3 Effect of training parameters on performance

For categorical cross-entropy, mean absolute error, mean squared error, and ordinal regression, the mean 4-class κ was .660 (95% CI .657-.664), .649 (95% CI .644-.653), .654 (95% CI .646-.661), and .664 (95% CI .659-.669), respectively. The performance of categorical cross-entropy and ordinal regression was significantly greater than mean absolute error (p = .011 and p = .004, respectively). The mean 4-class κ with no augmentation was .658 (95% CI .646-.666), compared to with augmentation .660 (95% CI .657-.664) (p = .675). The mean 4-class κ with random and equal class sampling at each mini-batch was .665 (95% CI .662-.669) and .660 (95% CI .657-.664), respectively (p = .135). For random class sampling, the predicted distribution of labels on the test set was 8.1% fatty, 47.5% scattered, 40.1% heterogeneously dense, and 4.3% extremely dense. This differed from the predicted distribution of labels on the test set with equal class sampling, which was 13.5% fatty, 37.5% scattered, 36.8% heterogeneously dense, and 12.2% extremely dense (p < .001, Fig. 4-7). The predicted binary distribution for random (44.4% dense) and equal sampling (49.0% dense) also differed (p < .001). For random class sampling, the mean sensitivity and specificity for classifying dense breast was .833 (95% CI .803-.856) and .888 (95% CI .872-.905), respectively. Comparatively, for equal class sampling, there was an increase in sensitivity (.880, 95% CI .869-.890, p < .05) with a decrease in specificity (.842, 95%

CI .828-.857, p < .001). A display of the range of diagnoses for models trained with different model and training parameters for 50 patients in the testing set is shown in Fig. 4-8.

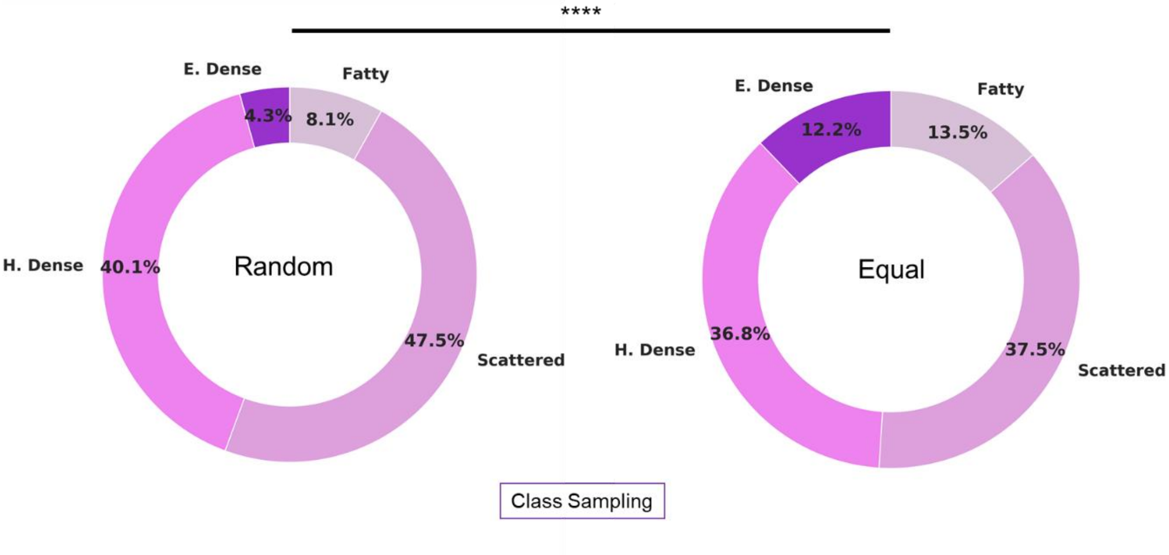


Figure 4-7. The distribution of predicted breast density labels in the testing set differed for experiments with random class sampling (left) compared to equal class sampling (right) at each mini-batch. P-values are denoted by ****p < .001

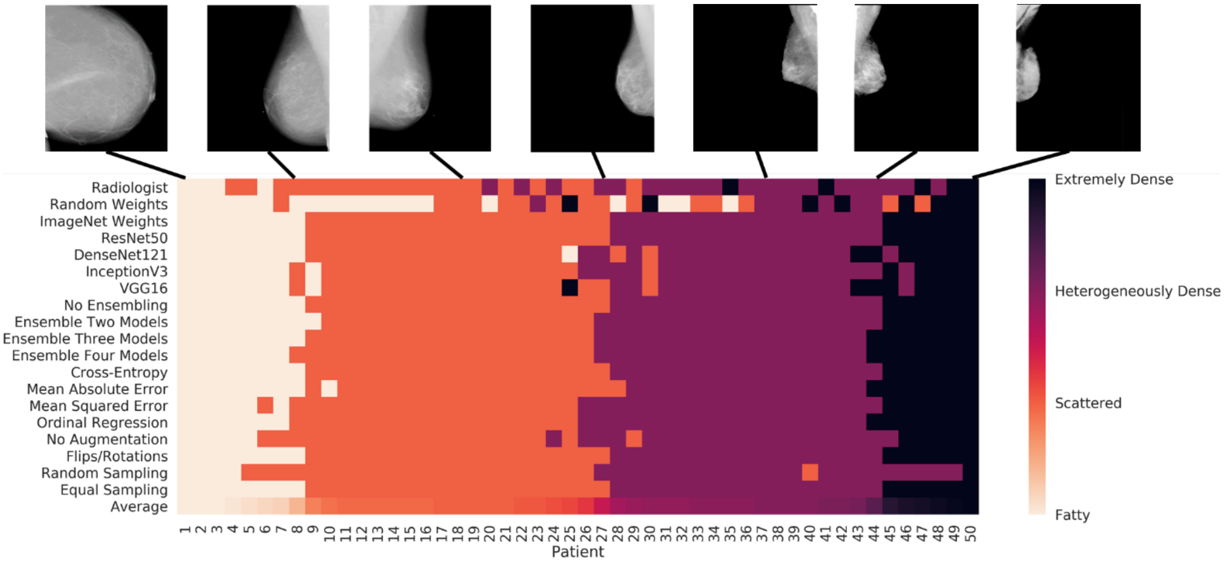


Figure 4-8. A visual display of the range of classifications for models trained with different model and training parameters for 50 patients in the testing set. The radiologist interpretation is displayed in the first row. The average breast density rating across all models and radiologist interpretation is displayed in the last row and was used to order the patients from least dense (left) to most dense (right).

For categorical cross-entropy, mean absolute error, mean squared error, and ordinal regression, the mean 2-class κ was .718 (95% CI .712-.724), .717 (95% CI .713-.721), .713 (95% CI .710-.716), and .721 (95% CI .718-.724), respectively. The performance of regression was significantly greater than mean absolute error (p = .009). The mean 2-class κ with no augmentation was .713 (95% CI .707-.718), compared to with augmentation .718 (95% CI .712-.724) (p = .287). The mean 2-class κ with random and equal class sampling at each mini-batch was .723 (95% CI .714-.729) and .718 (95% CI .712-.724), respectively (p = .471).

4.4.4 Contribution of breast size to density assessment

It has previously been shown that breast size is related to breast density. Specifically, smaller breast size is correlated with higher breast density.[219] Indeed, in our patient cohort, mammographic breast size was negatively associated with both radiologist and deep learning model determined breast density (Fig. 4-9). In looking at the AUC of ROC analysis for breast size vs density class, all values were above .5 (Fig. 4-10). Notably, AUC values were higher for the deep learning model than for the radiologist, suggesting that breast size was more important for the deep learning model than for the radiologist in determining breast density. It is also interesting to note that all AUC values were below .8, indicating that breast size is not the only factor that radiologist and the deep learning model takes into consideration in making the breast density assessment. Put in other words, the radiographic density within the breast is also an important factor. This serves as an important sanity check into what the deep learning model is learning as it is not just learning the trivial association between breast size and density.

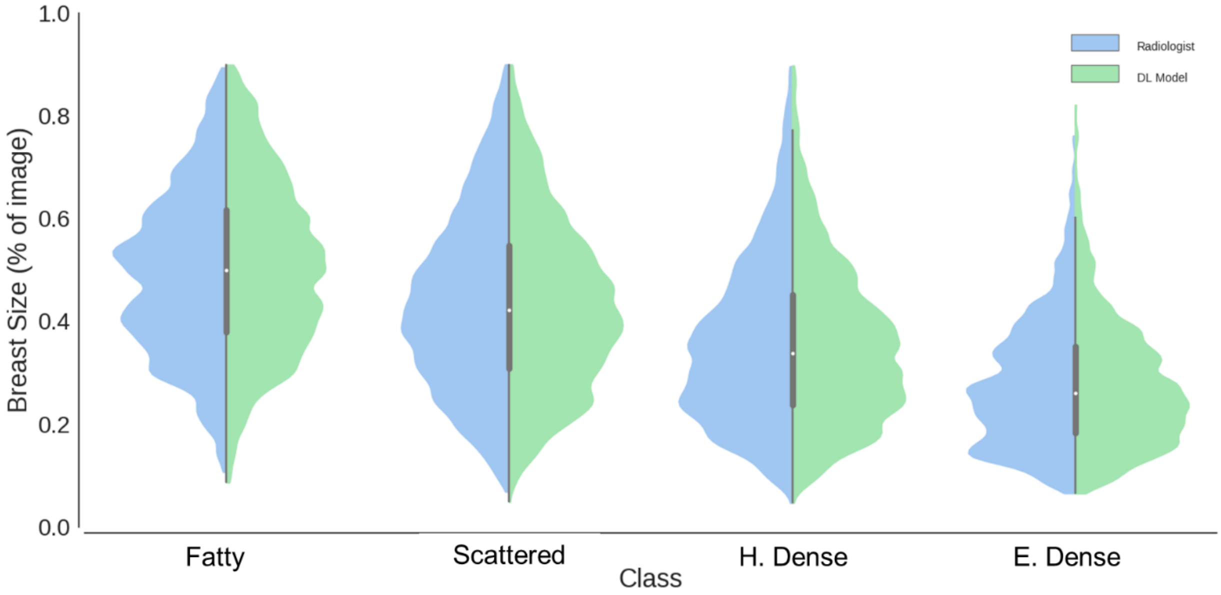


Figure 4-9. A plot of the distribution of breast size (% of mammogram) vs the radiologist/deep learning model determined breast density.

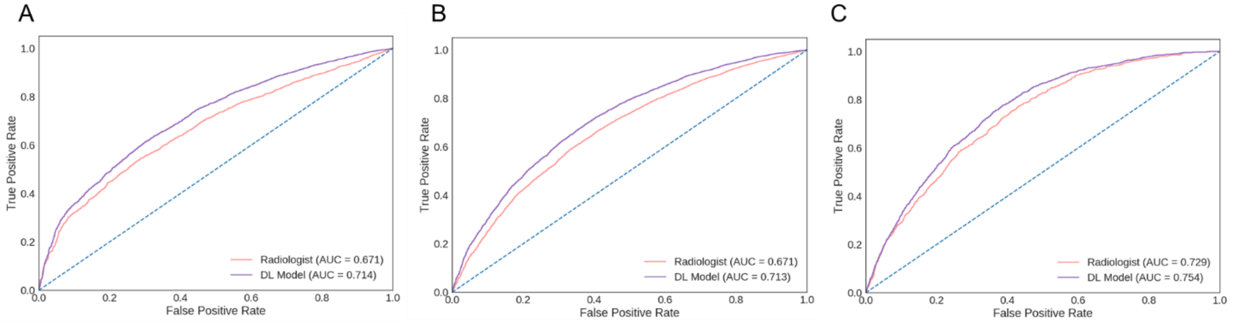


Figure 4-10. ROC of breast size (% of mammogram) vs the radiologist/deep learning model determined breast density for (A) fatty vs. scattered/heterogeneously dense/extremely dense, (B) fatty/scatter vs. heterogeneously dense/extremely dense, and (C) fatty/scatter/heterogeneously dense vs. extremely dense.

4.4.4 Differential performance across different patient races

Given the DMIST trial enrolled patients from 33 different sites across the United States and Canada, the dataset contained many patient races. Within each race, there were different distributions of breast density (Fig. 4-11). Notably, black or African American patients had a higher proportion of fatty breasts while Asian Americans had a higher proportion of dense breasts, consist with what has been previously reported within the literature.[219] When the model was trained on all races, there was differential performance across different patient races in the test set. The model had higher agreement with radiologists for Asian and white women. The model had lower agreement with radiologist for black or African American, Hispanic or Latino, and other or unknown races (Fig. 4-12-13). Notably, this did not change when the model was trained on only white women (Fig. 4-12, 4-14).

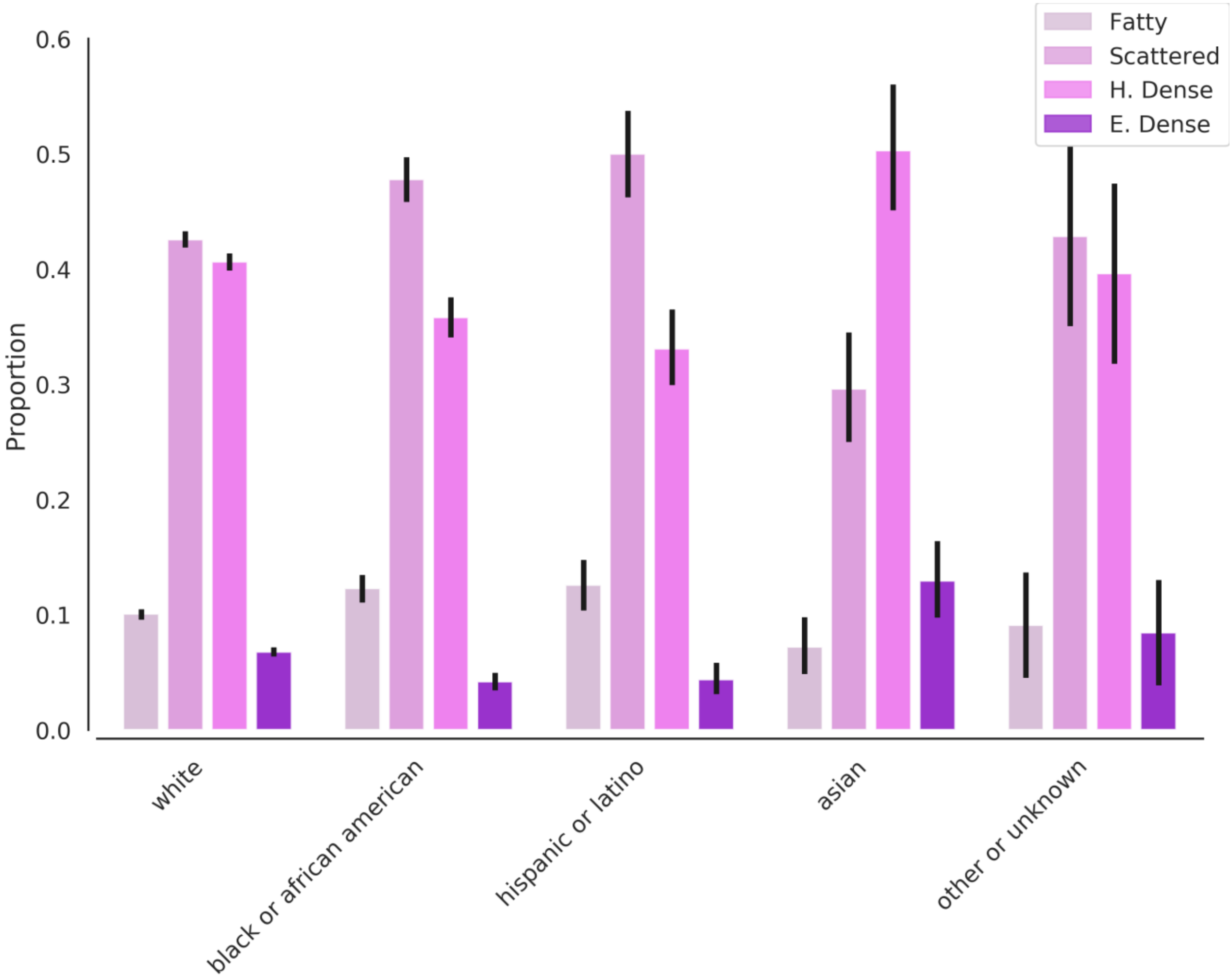


Figure 4-11. Distribution of different density classes vary across different races.

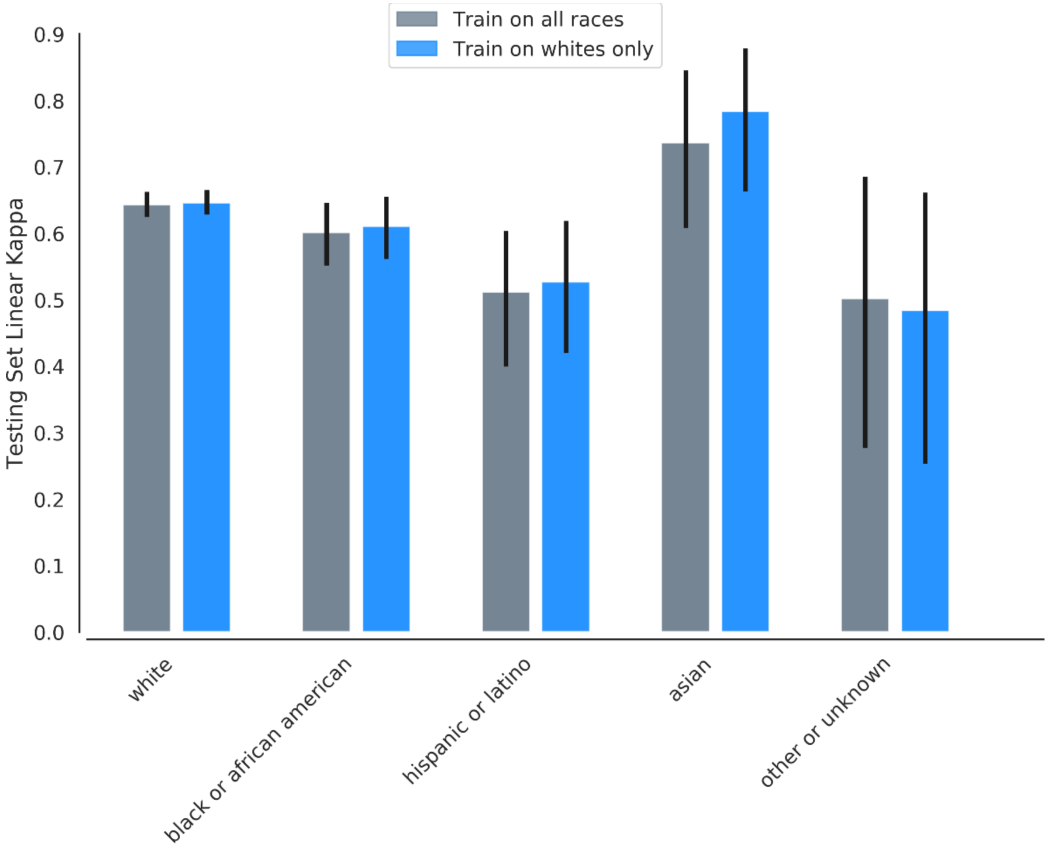


Figure 4-12. Agreement of a model trained on all races and a model trained on white women only with radiologists on the testing set, by race.

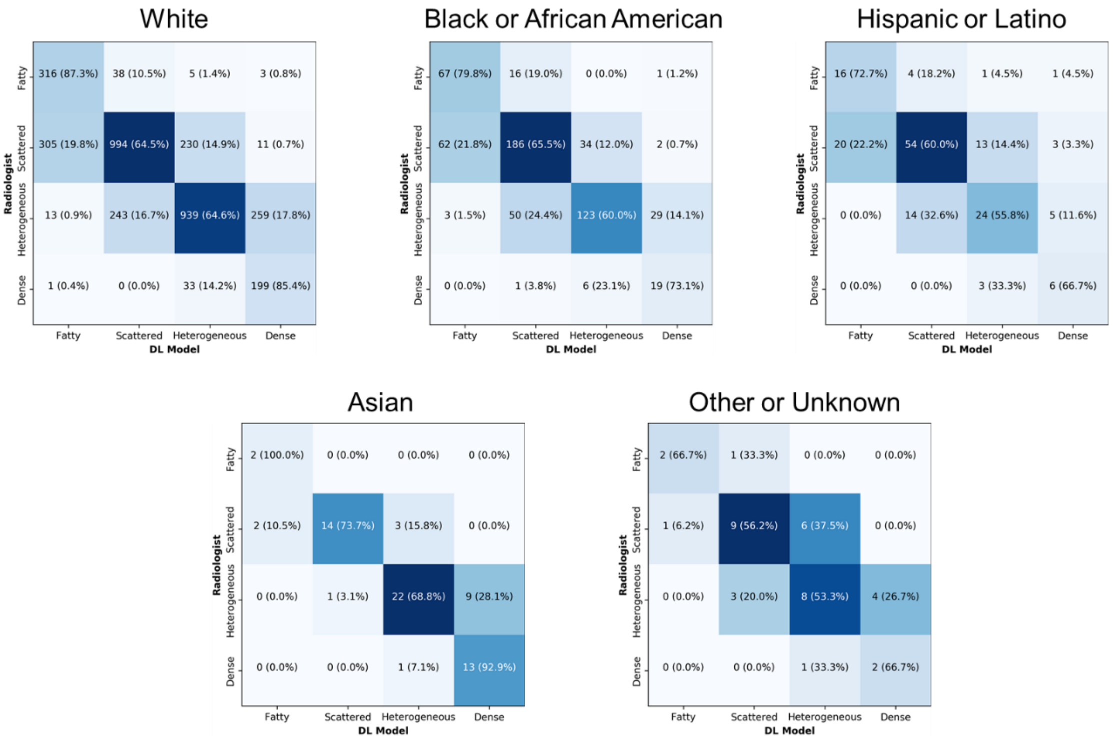


Figure 4-13. Confusion matrices between radiologist and a model trained on all races on the testing set, by race.

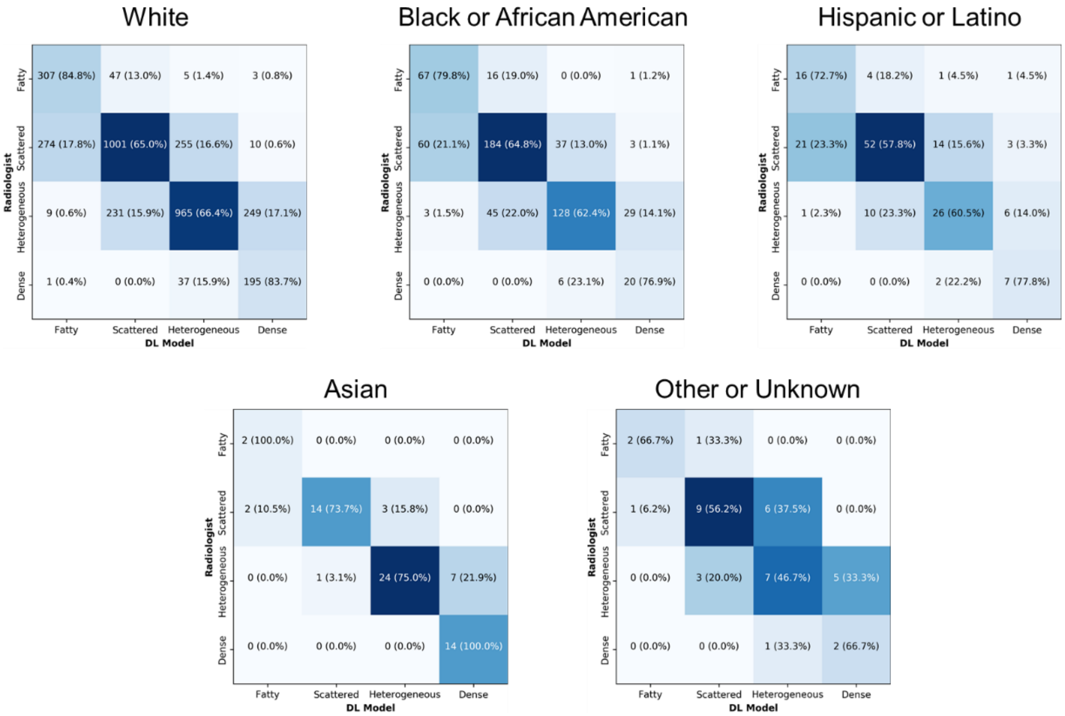


Figure 4-14. Confusion matrices between radiologist and a model trained on white women only on the testing set, by race.

4.4.5 Effect of digital-mammography data format on model generalizability

A plot of projections of intensity distributions of preprocessed images showed clustering within image format, delineating differences between image formats (Fig. 4-15A). Clustering by intensity distribution was preserved even after passing the images through a trained neural network, as shown by projections of the output of the penultimate layer, with the grouping by breast density occurring within the respective image format cluster (Fig. 4-16). For all image format specific models, testing set performance was decreased for other image formats compared to the image format the model was trained on (p < .001). In contrast, a model trained on all images showed no differences in performance across image formats (p > .05, Fig. 4-5B).

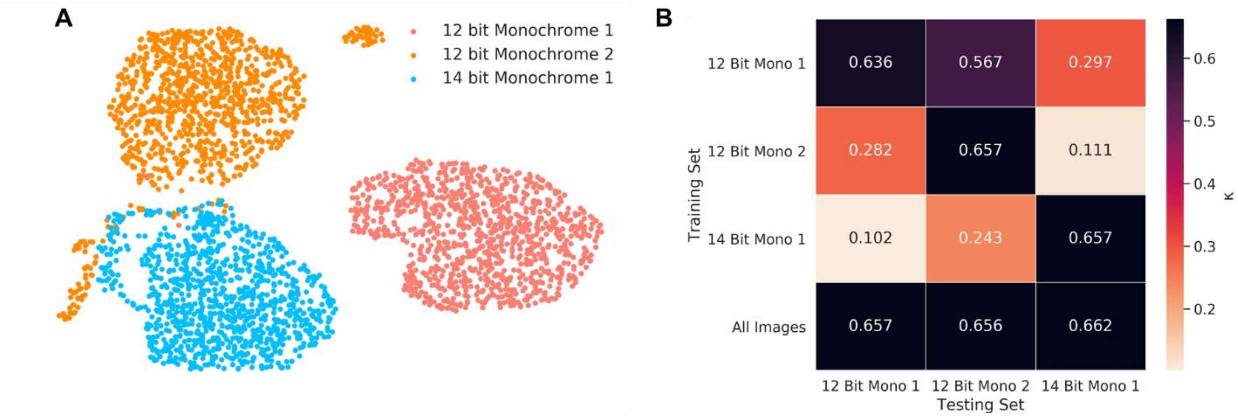


Figure 4-15. (A) Visualization of the histogram of intensities of 3000 preprocessed images from the testing set demonstrating clustering of images by image format. (B) Performance of models trained on specific image formats as well as all images, showing that for image format specific models, testing set performance was decreased for other image formats compared to the image format the model was trained on.

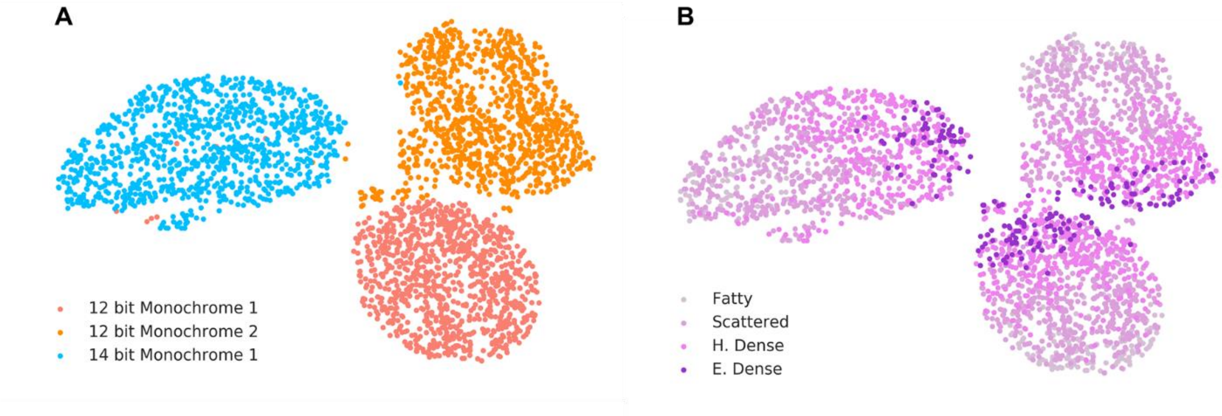


Figure 4-16. Visualization of an intermediate layer of the trained neural network for 3000 images in the testing set, color-coded by (A) image format and (B) radiologist interpretation of breast density.

4.4.7 Fine-tuning the model on a new patient cohort

We also acquired another patient cohort of 17,549 images from MGH. When a model trained on DMIST was applied to the MGH cohort, low performance was observed. When the model was fine-tuned on MGH, performance at MGH increased, but performance on DMIST decreased. Only when a model was trained on DMIST and MGH simultaneously was high performance on both DMIST and MGH observed (Fig. 4-17).

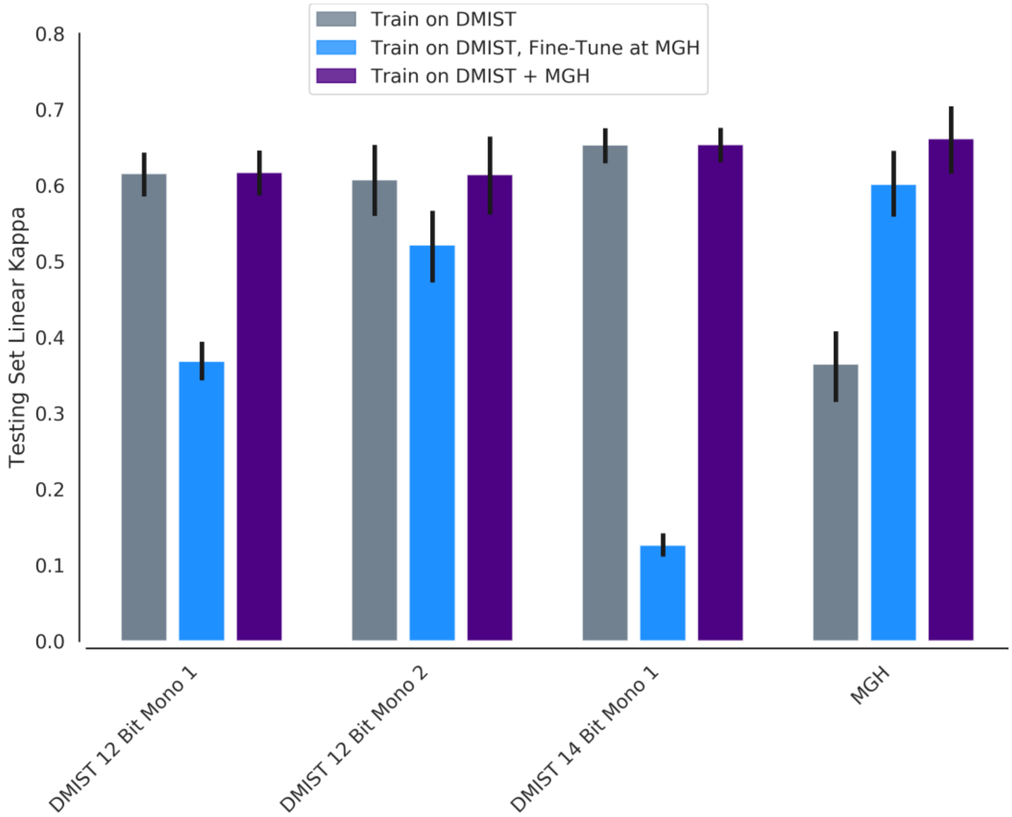
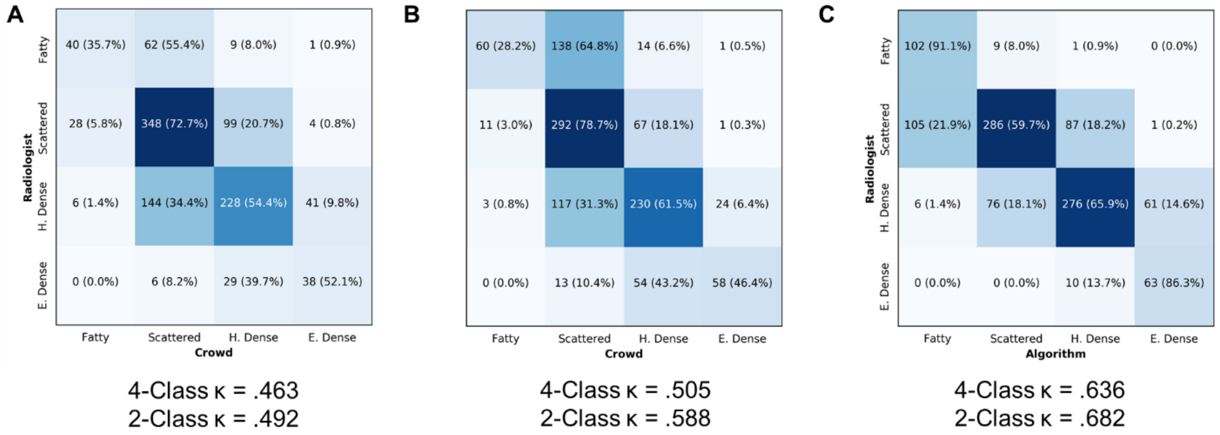


Figure 4-17. Testing set performance of a model trained on DMIST, trained on DMIST and then fine-tuned at MGH, and trained on DMIST and MGH simultaneously.

4.4.8 Crowdsourcing assessment

The 4-class κ between the crowd and algorithm (.505, 95% CI .503-.506) was greater than agreement between crowd and radiologist (.463, 95% CI .461-.464, p < .001, Fig. 4-18). Agreement with the algorithm was greater than agreement with the radiologist for both crowd participants who regularly read mammograms and those who do not (Fig. 4-19). Similarly, the 2-class κ between the crowd and algorithm (.588, 95% CI .587-.590) was greater than agreement between crowd and radiologist (.492, 95% CI .491-.495, p < .001). As a reference, the 4-class κ and 2-class κ between algorithm and radiologist was .636 (95% CI .635-.637) and .682 (95% CI .681-.684), respectively, for the same patient studies.



A

Radiologist \ Crowd	Fatty	Scattered	H. Dense	E. Dense
Fatty	40 (35.7%)	62 (55.4%)	9 (8.0%)	1 (0.9%)
Scattered	28 (5.8%)	348 (72.7%)	99 (20.7%)	4 (0.8%)
H. Dense	6 (1.4%)	144 (34.4%)	228 (54.4%)	41 (9.8%)
E. Dense	0 (0.0%)	6 (8.2%)	29 (39.7%)	38 (52.1%)

4-Class κ = .463
2-Class κ = .492

B

Crowd	Fatty	Scattered	H. Dense	E. Dense
	60 (28.2%)	138 (64.8%)	14 (6.6%)	1 (0.5%)
	11 (3.0%)	292 (78.7%)	67 (18.1%)	1 (0.3%)
	3 (0.8%)	117 (31.3%)	230 (61.5%)	24 (6.4%)
	0 (0.0%)	13 (10.4%)	54 (43.2%)	58 (46.4%)

4-Class κ = .505
2-Class κ = .588

C

Radiologist \ Algorithm	Fatty	Scattered	H. Dense	E. Dense
Fatty	102 (91.1%)	9 (8.0%)	1 (0.9%)	0 (0.0%)
Scattered	105 (21.9%)	286 (59.7%)	87 (18.2%)	1 (0.2%)
H. Dense	6 (1.4%)	76 (18.1%)	276 (65.9%)	61 (14.6%)
E. Dense	0 (0.0%)	0 (0.0%)	10 (13.7%)	63 (86.3%)

4-Class κ = .636
2-Class κ = .682

Figure 4-18. Confusion matrices showing the agreement between radiologist, algorithm, and crowd. The agreement between the algorithm and crowd (B) was greater than the agreement between crowd and radiologist (A). The agreement between algorithm and radiologist for the same patient studies (C) shown for reference.

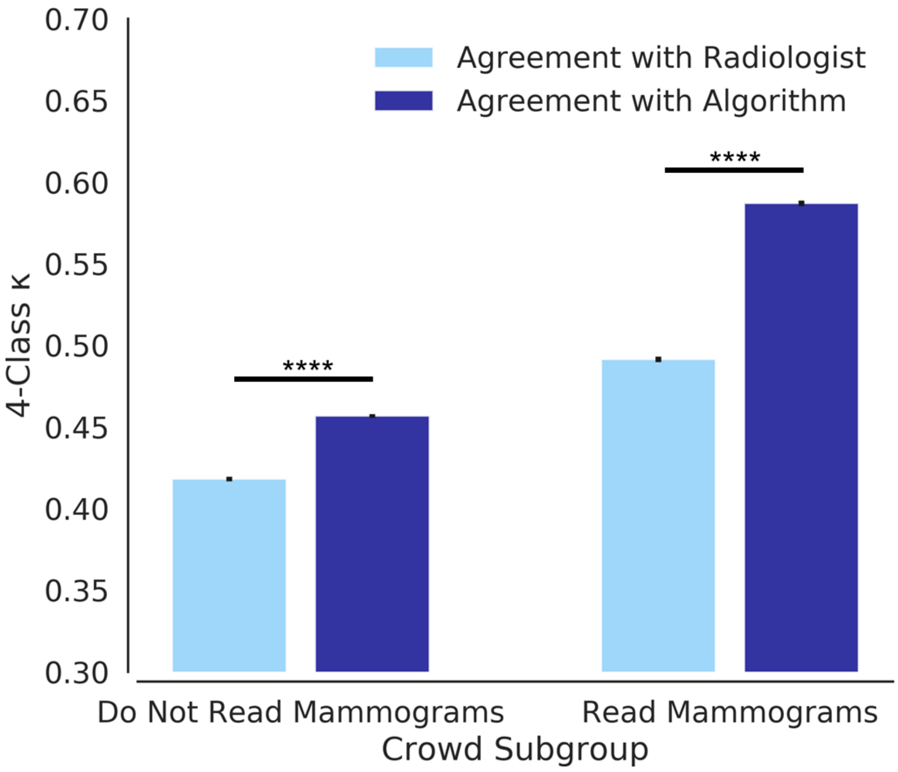


Figure 4-19. There was higher agreement, in terms of 4-class κ, with the algorithm than with the original interpreting radiologist from the DMIST trial for both crowdsourcing participants who read mammograms and those who do not. P-values are denoted by ****p < .001

4.5 Discussion

In this study, we investigated the performance of deep learning models in a large multi-institution and multi-mammography system patient cohort. Our best performing model achieved a κ of .667, equivalent to the agreement observed by Lehman et al., which only utilized mammograms from a single institution/mammography system.[202]

One challenge of training robust deep learning models is the availability of large annotated imaging datasets.[220] In this study, we provide empirical evidence that the size of the training set is a key determinant in the performance of neural networks, consistent with another study on abnormality classification in chest radiographs.[221] In accordance with deep learning studies in other domains, tens of thousands of annotated images are needed before model performance begins to plateau in diverse imaging cohorts, supporting the need for collaborative efforts among medical institutions.[5,37,221]

In our investigation of model parameters, pretraining and ensembling led to improvements in performance. Pretraining neural networks followed by fine-tuning in the target domain (also known as transfer learning) has become a well-established paradigm for medical imaging applications to achieve high performance.[37,40,222] The intuition behind this practice is that low level imaging features (such as texture, gradient, color, etc.) are similar across domains, allowing us to leverage large-scale datasets such as ImageNet, which has over 14 million annotated images, to learn these baseline filters learn task specific filters by training on the target data set. [220] In our study, we noted that pretraining on ImageNet improved performance for the breast density classification task. Further improvement in performance was seen with ensembling of independently trained models which is analogous to how a consensus of experts is more likely to be correct than any single expert.[223] Interestingly, neural network architecture did not have a significant effect on performance despite differences in model complexity and design.

One important consideration when training a model is the objective function used to optimize the algorithm, also known as a cost function. Our experiments have shown that the choice of cost function had a significant effect on model performance, mainly because each cost-function makes different assumptions about the nature of the labels. Specifically, mean absolute error, mean squared error, and ordinal regression assume that the categories are ordered while categorical cross-entropy does not. Furthermore, mean absolute error and mean squared error assume the distance between adjacent classes is equal whereas ordinal regression does not. In our application, breast density is classified on an ordered scale with undefined distances between adjacent classes (i.e. the distances between Fatty and Scattered compared to heterogeneously dense and extremely dense cannot be quantified), making ordinal regression the most appropriate cost function. This is validated in our experiments, where we find that ordinal regression exhibited the highest performance, although this was significantly different to only mean absolute error. We did not notice any effect of augmentation on performance. Augmentation is a commonly used approach to increase the diversity of the training data through random manipulations of the image.[1] While effective for some applications, we did not observe any statistically significant improvements from using data augmentation for our task when using the full training set. Augmentation may be of greater importance when less training data is available.

We also did not notice significant difference between random and equal class sampling on model performance in terms of κ coefficient. Class sampling is an important consideration in cases where there are differences in the number of patient samples from each class (i.e. when the majority class significantly outnumbers the minority class). In our study, we have more patients with scattered and heterogeneously dense breast (44.2% and 37.5% respectively) than with fatty and extremely dense breast (11.8% and 6.6%, respectively), which is the expected distribution as breast density has a normal distribution. Under random class sampling, the neural network would be exposed to more training examples of scattered and heterogeneously dense breast than of fatty and extremely dense breast. Equal class sampling can be used to mitigate this inherent class imbalance by ensuring that the neural network is adequately exposed to all classes. [224] However, it is also important to note that with equal class sampling, the distribution of predicted labels changes – specifically, minority classes are predicted with higher frequency and majority classes are predicted with lower frequency, as shown by our experimental results. The net result of this is that the sensitivity of predicting dense breast improves with equal class sampling. Moreover, equal class sampling leads to lower specificity for classification of dense breast. From a policy perspective, this can lead to more patients being notified that they have dense breast. If additional imaging is performed on these patients, this may lead to increases in the number of false positives. This is a key example of how the manner in which deep learning models are trained can have implications for clinical care.

It was observed that a model trained on all races performed agreed with radiologist to different degrees, depending on race. Notably, this did not change when the model was only trained on whites. This suggests that the differential performance across races is not due to poor generalization across imaging from patients of difference races, but rather differences in consistency of radiologist assessment for different races. Whether this is due to differing distributions of breast density across races (i.e. a radiologist is less consistent in calling certain classes of breast density) or other differences across races remains an open question.

One critical hurdle that prevents the deployment of deep learning models in the clinical work environment is their relatively poor generalizability across institutional differences, such as patient demographics, disease prevalence, scanners, and acquisition settings. In fact, other deep learning studies that have shown poor generalizability of deep learning models when applied to data from different institutions than the one they were trained on.[78,79] In our study, we found that models trained on specific digital-mammography data formats do not generalize to other data formats, and it was only after training on images from all digital-mammography data formats did our model achieve generalizability. Indeed, several deep learning studies for mammographic breast density assessment were only validated on patient cohorts from a single institution and/or digital-mammography systems.[202,225,226] Some possible differences between different digital-mammography systems or versions of systems include the x-ray tube target, filter, digital detector technology, and control of automatic exposure.[227] Our results add to the growing body of literature that states that deep learning models do not necessarily generalize when applied to data that differs from that which the model was trained with.

One approach to improve model performance at a new institution is to fine-tune the model at that institution. However, it was observed that this resulted in lower performance on the original training data. This phenomenon, known as catastrophic forgetting, can have implications for model deployment. Specifically, if a model that was FDA approved is subsequently fine-tuned, the model is now invalid for the original patient population that was used for evaluation. Regulations need to be put into place to ensure the integrity of deep learning models.

Alternatively, there are some approaches to mitigate catastrophic forgetting such as elastic weight consolidation and orthogonal weight modification.[171,172] Future studies can evaluate the effectiveness of these approaches for medical use cases.

Various studies have shown the utility of crowdsourcing and citizen science in biological and medical annotation.[228–230] As such, we performed a crowdsourcing assessment of our algorithm and radiologist annotations. Given the diversity of experience of the participants, it is unsurprising that the agreement between both the crowd and radiologist and the crowd and algorithm was lower than the agreement between algorithm and radiologist. Interestingly, the crowd had higher agreement with the algorithm than the crowd did with the radiologist, which may reflect the consistency of the algorithm in its assessment compared to the various radiologists in our multi-institutional study.

4.6 Limitations

There are several limitations to our study. First, we only investigated a scenario where hyperparameter are manual tuned and did not compare performance with AutoML, where model architecture and hyperparameters are tuned automatically.[231] Also, we only had one radiologist, from a cohort of radiologists, perform interpretation for each patient study. Future studies will incorporate multiple readers for each patient study. In addition, for models initialized with random weights, we did not optimize training hyperparameters such as the learning rate schedule or the duration of training.[232] It is possible that optimization would improve the performance of the randomly initialized model, but in this study, we show the performance advantage of pretrained neural networks with minimal hyperparameter tuning. Additionally, we only ensembled models trained with the same data, architecture, and training parameters. Ensembling models with differing parameters may further improve model performance.[233] Furthermore, in our investigation of augmentation, we only explored random flips and rotations, though future work will explore other augmentation techniques such as intensity scaling and elastic deformations.[234] Also, our crowdsourcing assessment included participants with a wide range of expertise. Future studies can further utilize crowdsourcing for evaluation with only experienced breast imaging radiologists as participants. Lastly, we only assessed mammography in this study – the incorporation of MR can further improve the clinical utility for assessment of breast cancer risk and outcomes.[235,236]

4.7 Conclusions

We showcase the various data, training, and model parameters that can influence model performance. Furthermore, we found that model performance deteriorates when training and testing on different imaging data formats. In performing this study in tandem with the development of the ACR AI-LAB, we demonstrate important principles that radiologists and data scientists have to consider when training neural network models. Our hope is that users of the AI-LAB can use this study as an educational tool when utilizing the AI-LAB to train their own deep learning models.

5 Enhancing glioma workflows with deep learning

5.1 Potential for automated tools for glioma

The current clinical workflow for a patient with neurological symptoms suspicious of brain malignancy is to receive MR imaging. This is followed by an invasive biopsy for pathological assessment and evaluation of molecular markers. Based on the information revealed by radiology and pathology, the patient receives a treatment that is a combination of chemotherapy, radiation, and surgery. The patient is then evaluated for treatment response using the Response Assessment in Neuro-Oncology (RANO) criteria, which is based on imaging and clinical information.[237] Based on how well the patient is responding to treatment, the patient is either continued on the current treatment course or considered for alternative approaches (Fig. 5-1). Within this clinical workflow, there are many potential opportunities for automated tools. The ones that I will focus on in this chapter are: 1) Detection and delineation of tumor boundaries, 2) Treatment response assessment[42], 3) Non-invasive prediction of the Isocitrate Dehydrogenase (*IDH*) molecular marker.[12]

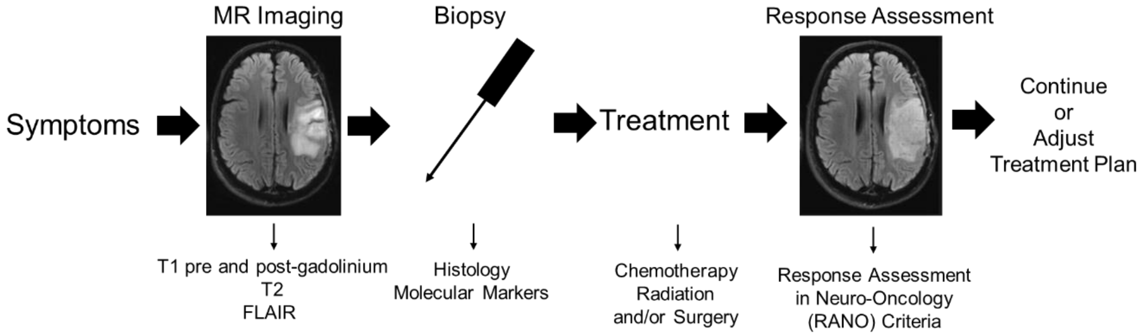


Figure 5-1. The current clinical workflow for patient with suspected glioma.

5.2 Background on automatic segmentation and response assessment

Gliomas are primary central nervous system (CNS) tumors with variable natural histories and prognoses depending on their histologic and molecular characteristics.[101] The current gold standards to determine treatment response and assess tumor progression in clinical trials are the Response Assessment in Neuro-Oncology (RANO) criteria.[89] For high-grade gliomas, including glioblastomas (GBMs), radiographic response assessment is based on 1) measuring the 2D product of maximum bi-dimensional diameters of contrast-enhancing tumor and 2) qualitative evaluation of T2/fluid-attenuated inversion recovery abnormal (FLAIR)-hyperintense regions.[89,95] However, manual delineation of tumor boundaries can be difficult due to the infiltrative nature of gliomas and presence of heterogeneous contrast enhancement, which is particularly common during anti-angiogenic treatment. As a result, there can be substantial inter-rater variability in 2D measurements for both contrast-enhancing and FLAIR-hyperintense tumors.[94,238,239] Furthermore, variability in segmentation can introduce substantial variability in calculated mean values of multi-parametric MR parameters, such as the volume transfer constant.[240] Consequently, there is great interest in developing reproducible automated methods for segmentation and calculation of the product of maximum bi-dimensional diameters. Although 2D linear measurements currently represent the gold standard for response assessment, volumetric measurements may capture tumor burden more accurately, particularly because gliomas are often irregularly shaped. However, volumetric response assessment has not been adopted for routine use due to the laborious efforts needed to perform tumor segmentation using existing tools and a lack of large-scale studies validating its benefit over simpler 2D approaches. A recent consensus paper on brain tumor imaging in clinical trials noted volumetric analysis as an improvement to current protocols.[241] An automated segmentation tool could help facilitate the use of tumor volume as a response endpoint in clinical trials and allow integration into the clinical work-flow. Rapid and reproducible tumor segmentation is also an essential step towards voxel-based quantitative assessment of single as well as multi-parametric imaging biomarkers of tumor response to treatment.[12,242–245]

With the advent of more powerful graphics processing units, deep learning has become the method of choice for automatic segmentation of medical images.[98,99] At the core of deep learning is the convolutional neural network; a machine learning technique that can be trained on raw image data to predict clinical outputs of interest. Existing deep learning methods have not been developed for the post-operative setting where the surgical cavity and brain distortion makes it difficult to reliably outline the boundaries of the tumor.[98,246]

There are two key challenges to automatic tumor segmentation. The first challenge is variability in brain extraction, an image pre-processing technique that separates the brain from skull and is essential for many neuroimaging applications.[247] Removing the skull from the image prevents automatic segmentation algorithms from falsely labeling non-brain regions as tumor and enables consistent intensity normalization across all patients. Many automated methods exist for brain extraction but their generalizability is limited [248–252]. Without manual correction, poor brain extraction can introduce errors in downstream automatic segmentation.[253] This is particularly important in the post-operative setting due to the widely heterogeneous and variable appearance of surgical cavity, calvarium, and scalp. The second challenge is generalizability: MR intensity values vary substantially depending on the MR scanner properties (including manufacturer, scanner type, and field strength) and acquisition parameters (including echo time, repetition time, and contrast injection dose/timing) and can result in substantial differences in tumor appearance.[241] Consequently, algorithms trained on limited datasets may not apply well to data acquired from different institutions, acquisition protocols, and patient populations.

In this section, we developed a fully automated pipeline for brain extraction and tumor segmentation that can be used to reliably generate abnormal FLAIR hyperintensity and contrast-enhancing tumor volumes as well as 2D bi-dimensional diameters according to the RANO criteria. Importantly, we utilize the biology in our approach – the brain encapsulates the FLAIR hyperintensity, which encapsulates the enhancing tumor. As such, we perform brain extraction first, followed by FLAIR hyperintensity segmentation which is fed as an input into the enhancing tumor segmentation (Fig. 5-2).[8]

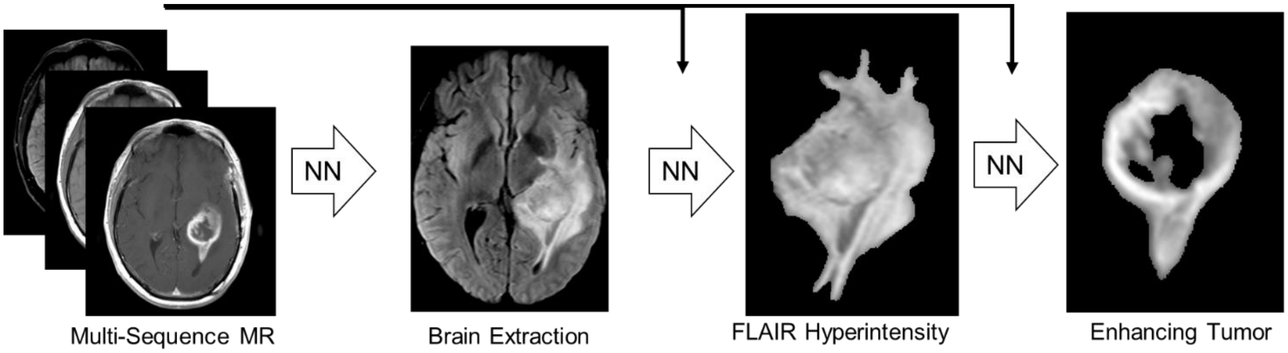


Figure 5-2. A sequential deep learning approach was designed for segmentation to utilize biological context.

We then validated the performance of the algorithm in both a multi-institutional pre-operative patient cohort and a longitudinal post-operative patient cohort from a single institution by comparing automated measurements to manual measurements derived from experts.

5.3 Materials and methods

5.3.1 Pre-operative patient cohort

The study was conducted following approval by the Hospital of the University of Pennsylvania (HUP) and the Partners Institutional Review Boards. Glioma patients at HUP, The Cancer Imaging Archive (TCIA), Massachusetts General Hospital (MGH), and Brigham and Women's Hospital (BWH) were retrospectively identified. The imaging study dates for HUP, MGH, and BWH ranged from 1998-2016. For the TCIA cohort, we identified glioma patients with pre-operative MRI data from TCGA and IvyGap.[254] All patients met the following criteria: (i) histopathologically confirmed grade II-IV glioma according to World Health Organization (WHO) criteria (2007 or 2016 criteria depending on whether the case occurred before or after 2016) and (ii) available preoperative MRI consisting of T2-weighted fluid attenuation inversion recovery (FLAIR) and post-contrast T1-weighted (T1 post-contrast) images. Patients were excluded if glioma was not histopathologically confirmed, either FLAIR or T1 post-contrast imaging was unavailable, or if there was excessive motion artifact on imaging. Demographics are shown in Table 5-1. The acquisition settings of the imaging for the pre-operative patient cohort are shown in Fig. 5-1 and 5-2. For the pre-operative cases, both 2D and 3D T1-weighted images were used, depending on which were available. 3D T1-weighted imaging was available for 29% of the patients in the pre-operative patient cohort. Our final pre-operative patient cohort included 239 patients from HUP, 293 patients from TCIA, 154 patients from MGH, and 157 patients from BWH.

	HUP, n = 239	TCIA, n = 293	MGH , n = 154	BWH, n = 157	Post-Operative, n=54
Gender (% Male)	54%	60%	60%	57%	61%
Age (yr)	53 (18-88)	54 (14-84)	52 (22-86)	48 (18-85)	65 (22-77)

Grade					
II	59	46	19	31	0
III	83	57	56	46	0
IV	97	190	79	80	54

Table 5-1. Age, gender, and histologic grade for the pre-operative (HUP, TCIA, MGH, and BWH) and post-operative patient cohorts (MGH). Note.- Age is shown as mean (minimum-maximum).

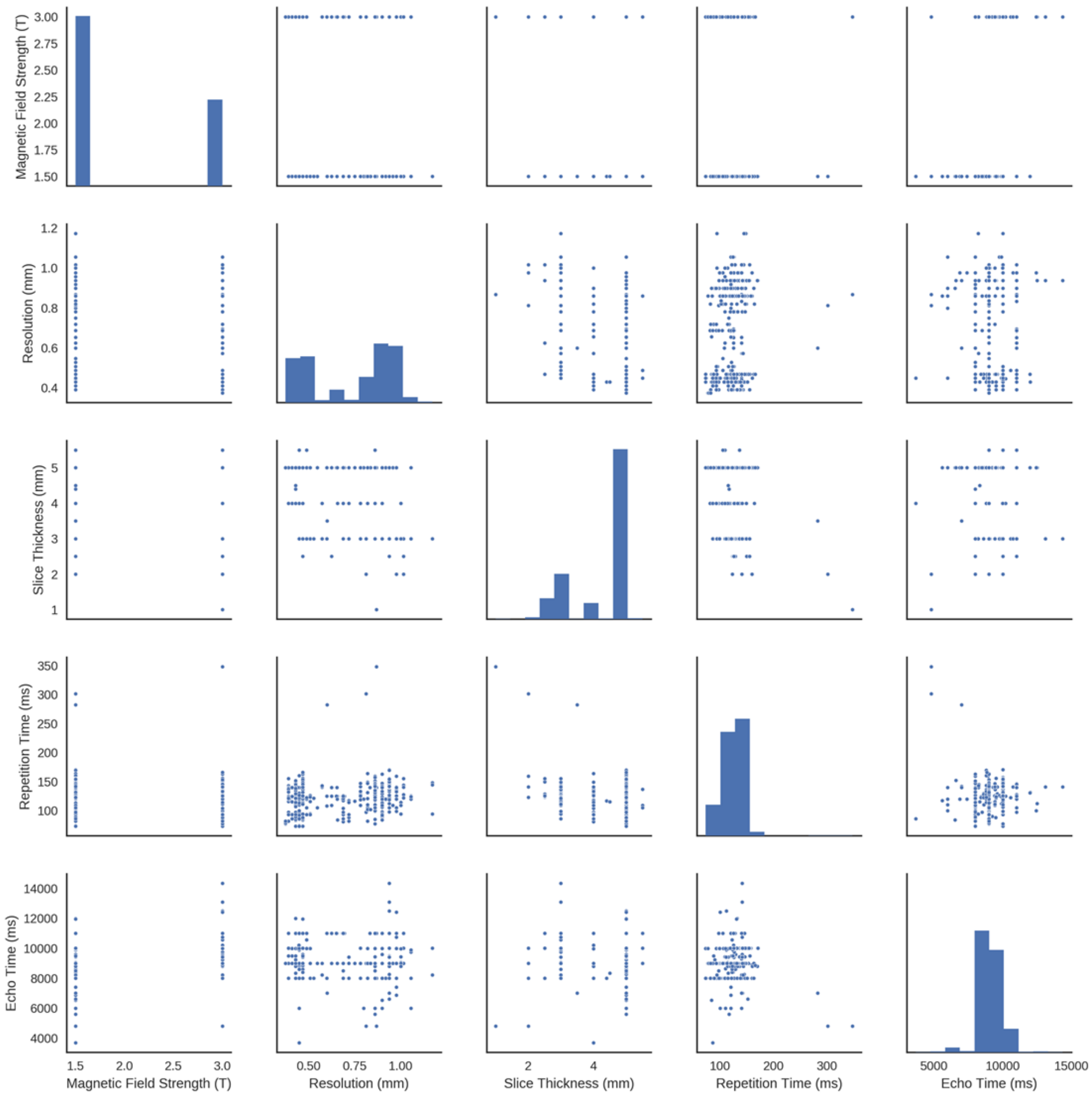


Figure 5-3. Scatter plots of Magnetic Field Strength, Resolution, Slice Thickness, Repetition Time, and Echo Time of FLAIR imaging in the pre-operative patient cohort. Histogram of frequencies are shown along the diagonal.

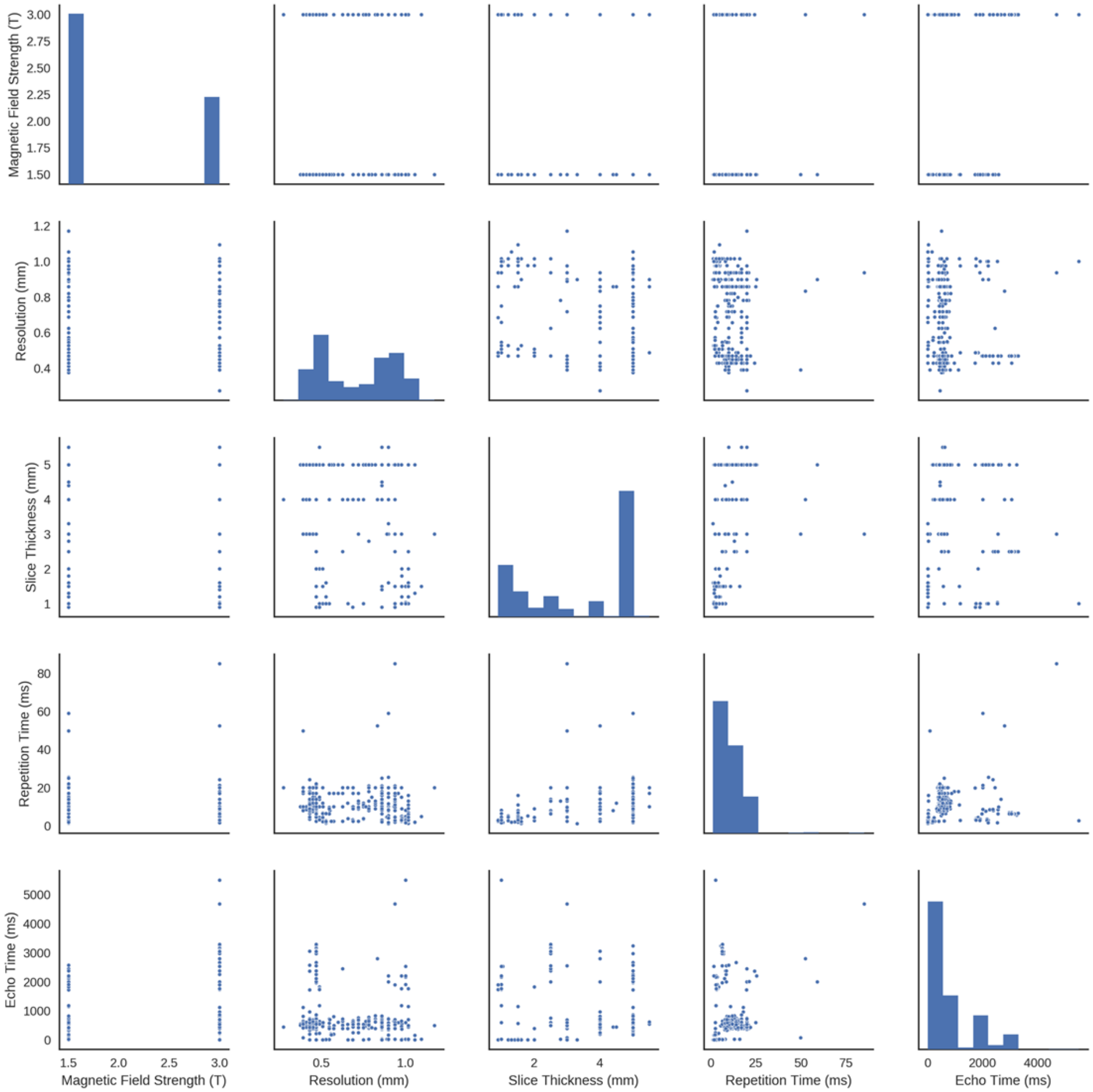


Figure 5-4. Scatter plots of Magnetic Field Strength, Resolution, Slice Thickness, Repetition Time, and Echo Time of T1 post-contrast imaging in the pre-operative patient cohort. Histogram of frequencies are shown along the diagonal.

5.3.2. Post-operative patient cohort

MRI data were acquired from two clinical trials at MGH that enrolled patients with newly diagnosed glioblastoma receiving standard chemoradiation (NCT00756106) or standard chemoradiation with cediranib (NCT00662506). There were 54 total patients. The Dana-Farber/Harvard Cancer Center IRB approved these studies. Inclusion criteria for both trials were age > 18 years, post-surgical residual contrast-enhancing tumor size of ≥ 1 cm in one dimension, histologically confirmed diagnosis of glioblastoma, and eligibility for standard therapy after surgery. For NCT00756106, MRI was performed at the following time points: within 1 week of starting chemoradiation therapy (baseline visit 1), 1 day before starting chemoradiation therapy (baseline visit 2), weekly during chemoradiation, and monthly before each cycle of adjuvant temozolomide until disease progression or at least until completion of six cycles of adjuvant temozolomide (whichever one occurred first).[255] MRI time points for NCT00662506 were previously described in Batchelor et. al.[256] MRI was performed at 3.0T (TIM Trio, Siemens Healthcare, Erlangen, Germany) and included FLAIR (TR = 10,000 ms, TE = 70 ms, 5-mm slice thickness, 1-mm inter slice gap, 0.43-mm in-plane resolution), and both pre- and post-contrast T1-weighted (TR = 600 ms, TE = 12 ms, 5-mm slice thickness, 1-mm interslice gap, 0.43-mm in-plane resolution) images. Our final post-operative patient cohort consisted of 713 visits from 54 patients from MGH. Twenty-one patient visits were excluded due to missing MRI sequences or excessive motion artifact.

5.3.3 Expert brain extraction, tumor segmentation, and RANO measurements

Brain extraction was performed in 42 randomly selected patients from the pre-operative and post-operative patient cohort by one rater (R.Y.H., neuroradiologist, 9 years experience). Manual tumor segmentations were performed on the FLAIR-hyperintense areas in the pre-operative patient cohort (Q.S., neuroradiologist, 5 years experience; R.Y.H.; A.B., neurosurgery resident, 5 years experience) and the FLAIR-hyperintense as well as contrast-enhancing tumor areas in the post-operative patient cohort (E.R.G., neuro-oncologist, 12 years experience, M.C.P., neuroradiologist, 11 years experience), with segmentation for each patient visit performed by a single expert evaluating both the pre- and post contrast MRIs to exclude post-operative blood products. Manual RANO bi-directional measurements as well as assessment for FLAIR progression were performed by two raters (E.R.G.; K.I.L., neuro-oncologist, 7 years experience) for both baseline visits, the visit with the lowest manual contrast-enhancing tumor volume, and the last patient visit from the post-operative patient cohort.[237]

5.3.4 Deep learning-based brain extraction

To remove any low-frequency intensity non-uniformity, N4 bias correction was applied to all MR images by subtracting the mean and then dividing by the standard deviation of the whole image (Fig. 5-5). Images were subsequently registered to FLAIR images using 3D Slicer.[257,258] The 42 patients for whom expert brain mask extraction was performed were divided into training (n = 30) and testing (n = 12) sets. The neural network was trained on the training set. As a point of reference, we compared brain extraction using our deep learning algorithm with that of other commonly used automatic brain extraction methods (Hybrid Watershed

Algorithm, Robust Learning-Based Brain Extraction, Brain Extraction Tool, 3dSkullStrip, and Brain Surface Extractor).[247–252] All methods were applied to the T1 post-contrast images using default parameters except for Robust Learning-Based Brain Extraction, which has no tunable parameters.

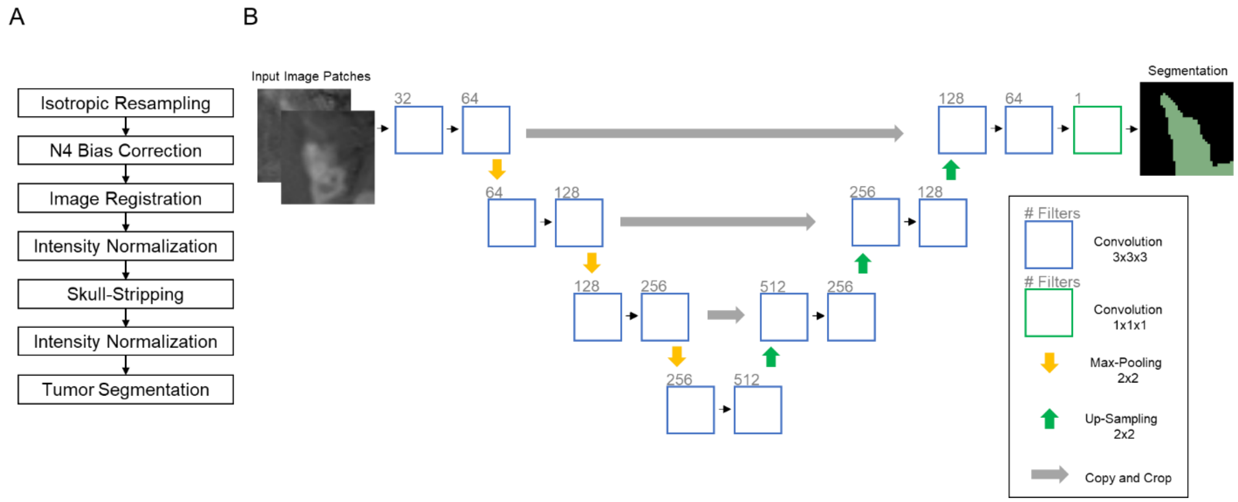


Figure 5-5. (A) Image pre-processing steps in our proposed approach. (B) A U-Net architecture was used for skull-stripping and tumor segmentation. The input is a patch from FLAIR, T1 pre-contrast, T1 post-contrast, and/or FLAIR tumor region depending on the segmentation task. The output is a binary label map.

5.3.5 Deep learning-based abnormal FLAIR hyperintensity and contrast-enhancing Tumor Segmentation

Following brain extraction, N4 bias correction was re-applied to brain tissue only. In order to evaluate the performance of brain extraction and segmentation separately, any tumor region, that was identified by the expert manual segmentation and removed during our brain extraction method was reinserted.

The HUP, TCIA, and MGH patient pre-operative cohorts were randomly divided into training and testing sets in a 4:1 ratio. The BWH patient cohort was used as an independent testing set. A single neural network model was trained for FLAIR hyperintensity segmentation in the pre-operative patient cohort using only the training set. Once the model was trained, performance was assessed on the testing and independent testing sets.

The patients from the single institutional post-operative patient cohort were randomly divided into training and testing sets in a 4:1 ratio. Data were split on a patient level such that all visits for a patient were either entirely in the training or test set (Fig. 5-6). Two neural network models were trained for the post-operative patient cohort: FLAIR hyperintensity segmentation and contrast-enhancing tumor segmentation. Only the training set was used during training of the model. Once trained, the performance of the model was assessed on the separate testing set.

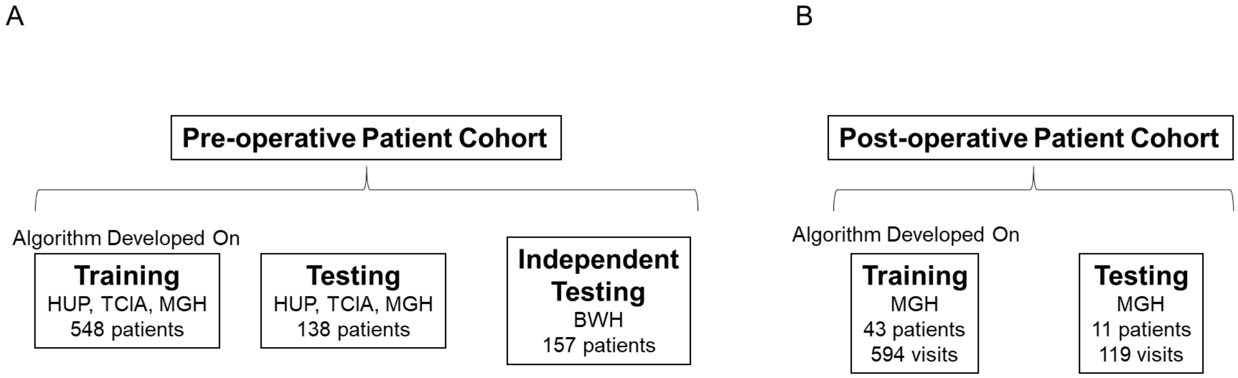


Figure 5-6. (A) Division of Training, Testing, and Independent Testing Sets in the preoperative patient cohort. (B) Division of Training and Testing in the post-operative patient cohort.

5.3.6 Neural network architecture and post-processing

We utilized the 3D U-Net architecture, a neural network designed for fast and precise segmentation, for both brain extraction and tumor segmentation (Fig. 5-5B).[8,259] Similar to the original 2D U-Net, our architecture consists of a downsampling and an upsampling arm with residual connections between the two that concatenate feature maps at different spatial scales. The networks were designed to receive input patches from multiple channels: 1) FLAIR and T1 post-contrast images for brain extraction, 2) FLAIR and T1 post-contrast images for FLAIR hyperintensity segmentation in the pre-operative patient cohort, 3) FLAIR, T1 pre-contrast, and T1 post-contrast images for FLAIR tumor segmentation in the post-operative patient cohort, and 4) FLAIR, T1 pre-contrast, T1 post-contrast, and FLAIR hyperintensity region for contrast-enhancing tumor segmentation in the post-operative patient cohort. Rectified linear unit activation (ReLU) was used in all layers, with the exception of the final sigmoid output. Batch normalization was applied after each convolutional layer for regularization. We used Nestorov Adaptive Moment Estimation to train the 3D U-Nets with an initial learning rate 10^{-5}, minimizing a soft Dice loss function:

$$(1)\ \ D(p,g) = \frac{2\sum_i g_i p_i}{\sum_i (g_i + p_i) + \alpha}$$

where D is Dice, p is the probability output of the neural network, g is the ground truth, and α is a constant. Our networks were implemented in DeepNeuro with Keras/Tensorflow backend.[10] Each U-Net was trained on a NVIDIA Tesla P100 graphics processing unit. During training, 20% of the training set was withheld as a validation set. For brain extraction, 50 patches (64x64x8) were extracted, randomly, for each patient in the training set and 10 patches were extracted for each patient in the validation set. For tumor segmentation, 20 patches (64x64x8) were extracted from normal brain and FLAIR hyperintense regions in a 1:1 ratio for each patient in the training set and 4 patches were extracted for each patient in the validation set. Before patches were used to train the network, they were augmented by means of sagittal flips.

Augmentation increases the size of the training set while also preventing overfitting.[12] The network was trained through all extracted patches until the validation loss did not improve for 10 consecutive iterations. Once the network was trained, inference was performed by gridding the MR images into patches at 8 different offsets from the upper-most corner of the image. The model then predicted probability maps for each of these patches, and voxels with predictions from multiple overlapping patches had their probabilities averaged. For prediction of the contrast-enhancing tumor regions, the output probability map from the FLAIR hyperintensity segmentation neural network was used as input instead of the manually derived FLAIR hyperintensity region.

5.3.7 AutoRANO algorithm

We developed an AutoRANO algorithm to automatically derive RANO measurements from our automatic deep-learning based contrast-enhancing tumor segmentations as described above. The algorithm searches for the axial slice with the largest tumor area and determines if the lesion is measurable. A measurable lesion was defined as a minimum length of both perpendicular measurements greater than or equal to 12 mm (based on a threshold of 10 mm if slice thickness + gap <= 5 mm or a threshold of 2 x [slice thickness + gap] if slice thickness + gap > 5 mm).[88] If the lesion was measurable, the product of maximum bi-dimensional diameters was automatically derived by first exhaustively searching for the longest diameter and then the corresponding longest perpendicular diameter. The angle between the longest diameter and the perpendicular diameter was restricted to 85-95º. If there was more than one measurable lesion on the same scan, the products of maximal bi-dimensional diameters were summed (for up to 5 measurable lesions).[88] The AutoRANO algorithm was applied to the automatically segmented contrast-enhancing tumor regions (Fig. 5-7C).

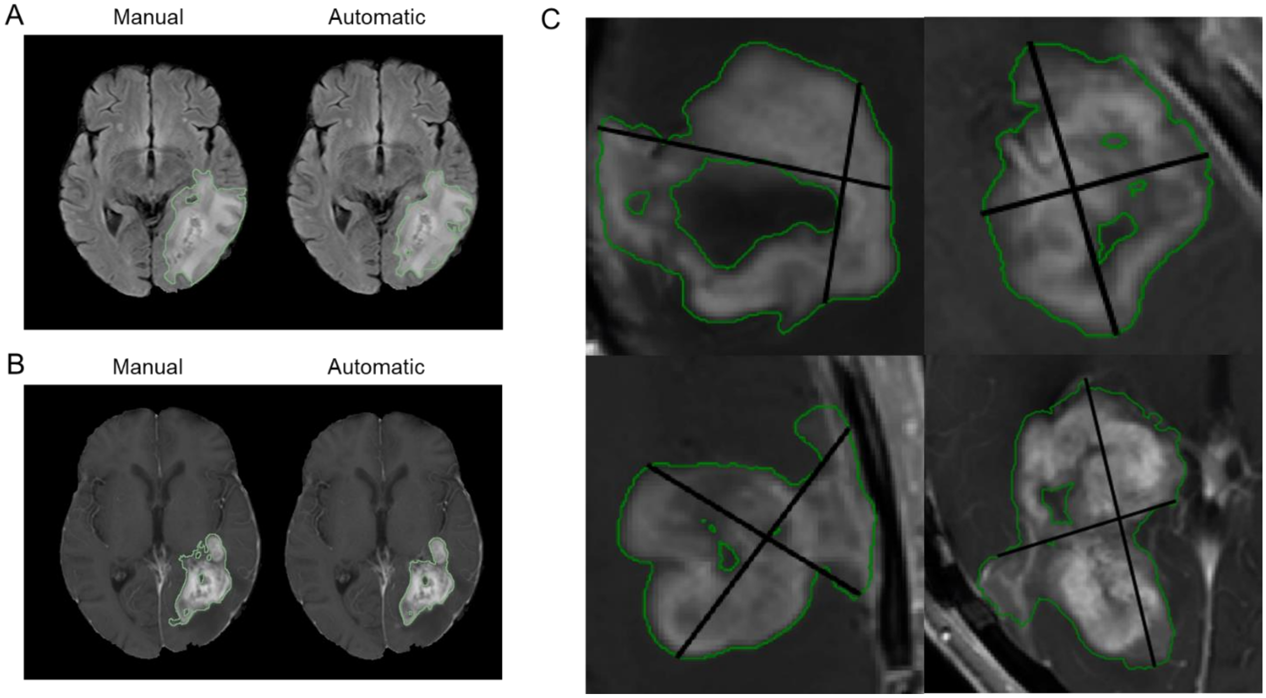


Figure 5-7. (A) Example of manual vs automatic FLAIR hypertintensity segmentation (A) and enhancing-tumor segmentation (B) for the testing set in the post-operative patient cohort. (C) Examples of AutoRANO applied to automatic enhancing segmentations on the post-operative patient cohort.

5.3.8 Comparison of manual vs automatic determination of response assessment

To compare manual vs automatic determination of no progression vs progression, the nadir of the baseline visits and the patient visit with the smallest manual contrast-enhancing tumor volume were compared with the last patient visit. For manual raters, progression was defined as ≥25% increase in RANO measurement, presence of FLAIR progression, or the presence of new measurable contrast-enhancing lesions.[88,237] Similar criteria were applied for AutoRANO with FLAIR progression defined as ≥40% increase in FLAIR hyperintensity volume. Agreement between Rater 1, Rater 2, and Auto RANO were assessed by means of unweighted κ coefficient.

5.3.9 Statistical analysis

Neural network segmentation was compared with expert segmentation by means of Sørensen–Dice coefficient, sensitivity, and specificity, and evaluated statistically using the Dunnet's test (significance level $p < 0.05$). Comparison of volume and RANO measurements were assessed via either the Spearman's rank correlation coefficient (ρ) or intraclass correlation coefficient (ICC) (significance level $p < 0.05$). For the assessment of rater-algorithm agreement for volumes in both the pre-operative and post-operative patient cohorts, where each patient visit has only one rater perform segmentation, ICC estimates were calculated based on a single measurement, absolute-agreement, one-way random-effects model. For assessment of intra-rater/algorithm reliability in the double baseline visit in the post-operative patient cohort, ICC estimates were calculated based on a single measurement, absolute-agreement, two-way mixed-effects model. For assessment of inter-rater and rater-algorithm agreement for RANO measures in the post-operative patient cohort, where each rater as well as the algorithm assessed the same patient visits, ICC estimates were calculated based on a single measurement, absolute-agreement, two-way random-effects model. R package IRR (inter rater reliability) was used for ICC computation.[260] For the post-operative patient cohort, the nadir was defined as the minimum volume or minimum 2D linear measurements at any time point from baseline to last visit. For longitudinal comparison of volume and RANO measurements, the last patient visit was assessed relative to the nadir (delta measure = volume or RANO measure of the last patient visit – volume or RANO measure of the nadir).

5.4 Results

5.4.1. Deep learning-based brain extraction

We compared brain extraction using our deep learning algorithm, based on the 3D U-Net architecture[259], with that of both human expert and commonly used brain extraction software packages. The mean Dice score between our algorithm and manual expert brain extraction was 0.935 (95% CI, 0.918-0.948) in the testing set (Supplementary Table 2, Fig. 5-8A). Compared to other commonly used brain extraction techniques (Table 5-2), our algorithm had the highest Dice score and specificity for the testing set. When the U-Net was applied to all 843 patients in the pre-operative patient cohort, the mean fraction of FLAIR hyperintensity retained in the extracted brain image (defined as tumor volume remaining in the brain-extracted image divided by total tumor volume) was 0.987 (95% CI, 0.984-0.990, Fig. 5-8B). When applied to the 713 patient visits in the post-operative patient cohort, the mean fraction of FLAIR hyperintensity and contrast-enhancing tumor retained in the extracted brain image was 0.996 (95% CI, 0.994-0.997, Fig. 5-8C) and 0.982 (95% CI 0.977-0.987), respectively.

	Dice		Sensitivity		Specificity	
	Training (n = 30)	Testing (n = 12)	Training (n = 30)	Testing (n = 12)	Training (n = 30)	Testing (n = 12)
U-Net	**0.954**	**0.935**	**0.966**	0.956	0.986	**0.978**
Hybrid Watershed Algorithm	0.905*	0.900	0.921	**0.978**	0.976	0.949
Robust Learning-Based Brain Extraction	0.925	0.911	0.897**	0.943	**0.986**	0.965
Brain Extraction Tool	0.826***	0.854*	0.954	0.966	0.909***	0.923*
3dSkullStrip	0.841***	0.848*	0.956	0.965	0.916***	0.919*
Brain Surface Extractor	0.812***	0.800***	0.740***	0.773***	0.983	0.961

Table 5-2. Mean Dice, sensitivity, and specificity compared to human expert brain extraction our deep learning algorithm (based on 3D U-net architecture) versus other commonly used skull-stripping methods within the training and testing sets. Note.- Methods with the highest performance are shown in bold. Dunnet's test was used to compare significance between 3D U-Net and the other skull-stripping methods. *p < 0.05, **p < 0.01, ***p < 0.001

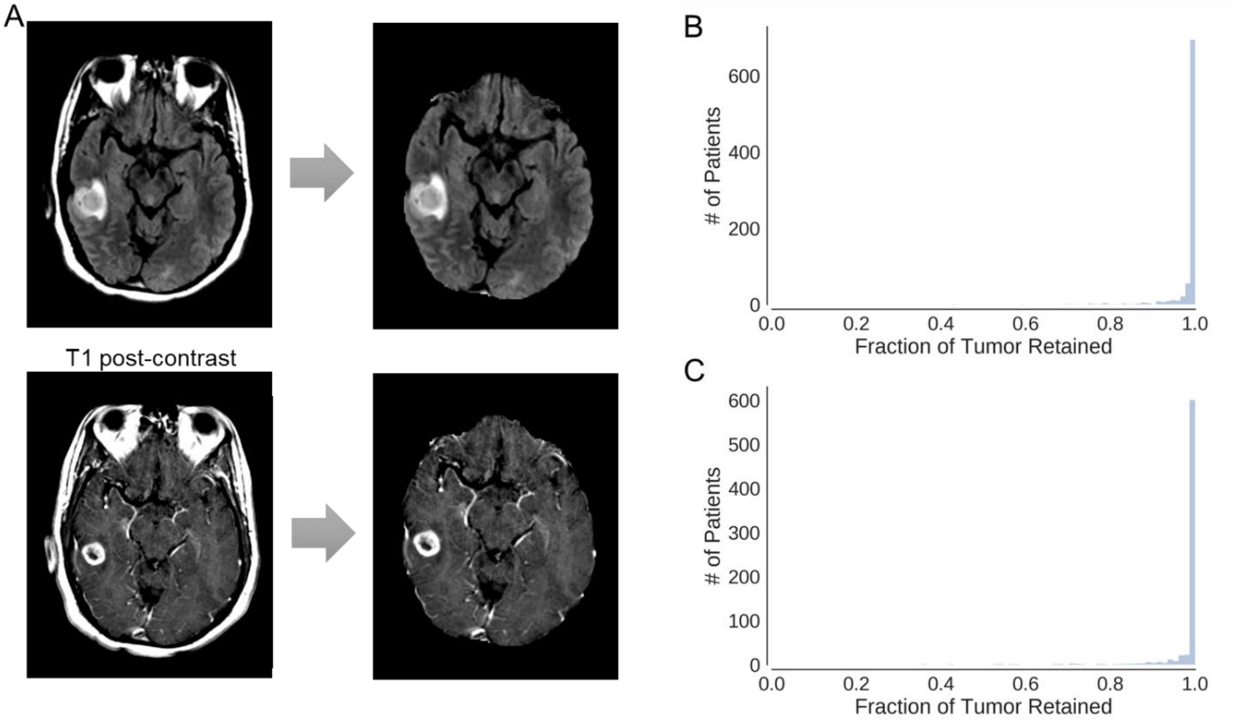


Figure 5-8. (A) Example of skull-stripping using 3D U-Net (B) Histogram plot of the fraction of FLAIR tumor that was retained after brain extraction for all patients in the pre-operative patient cohort (n = 843) (C) Histogram plot of the fraction of FLAIR hypertintensity that was retained after brain extraction for all patient visits in the post-operative patient cohort (n = 713)

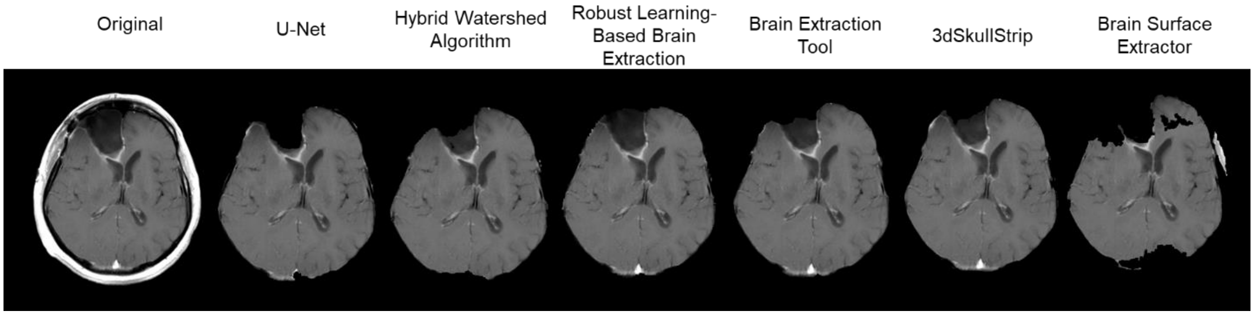


Figure 5-9. Example of the performance different brain extraction methods on a test patient visit from the post-operative patient cohort. This patient visit had a fluid-filled resection cavity, a clear surgical point of entry, and enhancing adhesions. In this scenario, the U-Net performed better than the other brain extraction methods at removing the resection cavity while not removing brain tissue.

5.4.2 Deep learning-based FLAIR hyperintensity and contrast-enhancing tumor volume segmentation

The average time for brain extraction, FLAIR hyperintensity, and contrast-enhancing tumor segmentation was 19 seconds using our trained algorithms. For the testing set of the pre-operative patient cohort, the mean Dice score for FLAIR hyperintensity segmentation was 0.796 (95% CI 0.753-0.803) (Table 5-3). For the independent testing set, the mean Dice score for automatic FLAIR hyperintensity segmentation compared to expert human segmentation was 0.819 (95% CI 0.793-0.842). Examples of FLAIR hyperintensity segmentations for the independent testing set of the pre-operative patient cohort are shown in Fig. 5-10. For the testing set of the post-operative patient cohort, the mean Dice score for automatic FLAIR hyperintensity segmentation compared to manual segmentation was 0.701 (95% CI 0.670-0.731). The mean Dice score for automatic segmentation compared to manual contrast-enhancing tumor segmentation was 0.696 (95% CI 0.660-0.728).

	HUP		TCIA		MGH		BWH		Total			
	n	Dice	n	Dice	n	Dice	n	Dice	n	Dice	ρ	ICC
Training	191	0.785	239	0.823	118	0.822			548	0.810	0.948	0.926
Testing	48	0.812	59	0.801	31	0.763			138	0.796	0.914	0.873
Independent Testing							157	0.829	157	0.819	0.957	0.923

Table 5-3. Mean Dice similarity coefficient for automatic versus expert manual segmentations of FLAIR tumor. Spearman's rank coefficient (ρ) and intraclass correlation coefficient (ICC) were calculated to show agreement in volumes derived from automatic and manual segmentations.

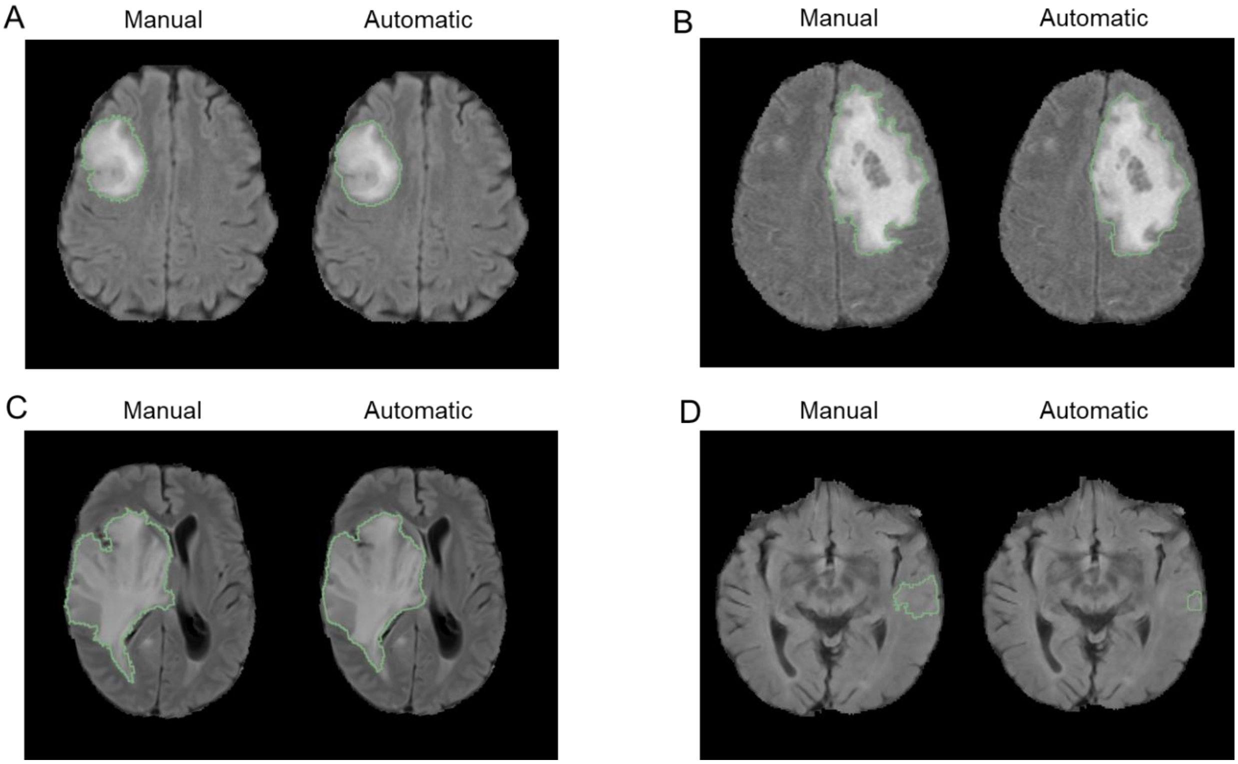


Figure 5-10. (A) Example of manual vs automatic FLAIR hypertintensity segmentation of a (A) grade II, (B) grade III, (C) grade IV glioma for the independent testing set of the pre-operative patient cohort. An example of automatic segmentation with poor agreement with expert manual segmentation (grade III) is shown in (D). Segmentations shown are overlaid on axial FLAIR image.

Examples of FLAIR hyperintensity and contrast-enhancing tumor segmentations for the testing set of the post-operative patient cohort are shown in Fig. 5-7A-B. Examples of longitudinal tracking of FLAIR hyperintensity and contrast-enhancing tumor volumes for two patients in the testing set are shown in Fig. 5-11. The ICC for calculated FLAIR hyperintensity volumes between automatic and manual segmentation was 0.915 ($p < 0.001$) in the pre-operative and 0.924 ($p < 0.001$) in the post-operative patient cohorts. The ICC for calculated FLAIR hyperintensity volumes between automatic and manual segmentation in the pre-operative patient cohort was 0.901 ($p < 0.001$) for Grade II, was 0.936 ($p < 0.001$) for Grade III, and was 0.899 ($p < 0.001$) for Grade IV. The ICC for contrast-enhancing tumor volume in the post-operative patient cohort was 0.965 ($p < 0.001$, Fig. 5-12). In the rare cases when the algorithm was off, the reason was due to similarity in signal intensity between normal brain and tumor – a similar challenge for human readers (Fig. 5-10D and Fig. 5-13).

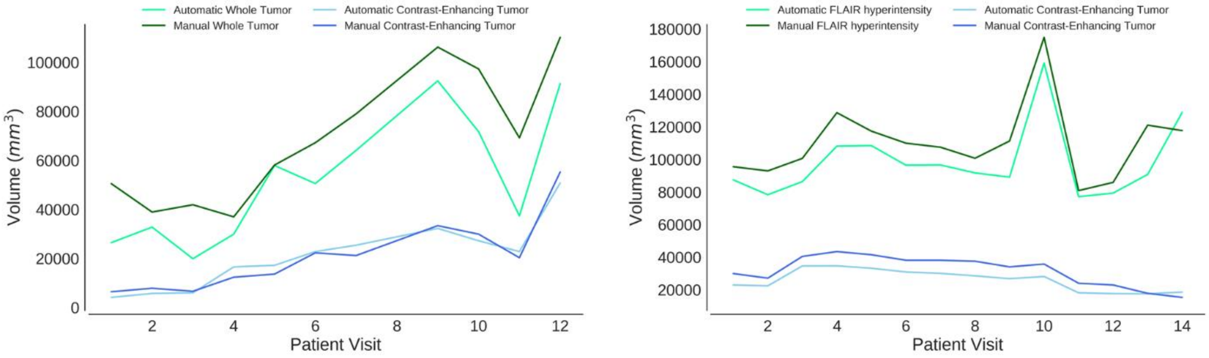


Figure 5-11. Longitudinal plots of automatic vs manual volumes for 2 patients in the testing set of the post-operative patient cohort, demonstrating that automatic measurements mirror the pattern of volume changes seen with manual measurements.

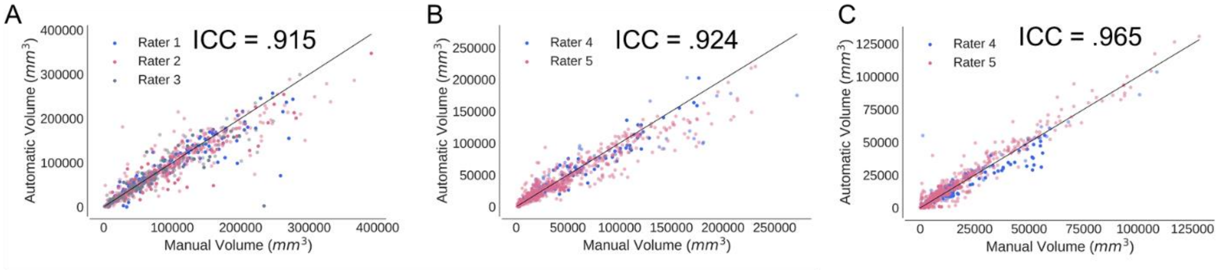


Figure 5-12. Automatically and manually derived volumes are highly correlated. Correlation between manually and automatically derived volumes for (A) FLAIR hypertintensity in the pre-operative patient cohort, (B) FLAIR hyperintensity in the post-operative patient cohort, and (C) contrast-enhancing tumor in the post-operative patient cohort. Training and Testing Sets are shown light blue/red/gray and dark blue/red/gray, respectively. Line of identity (x = y) is shown in all plots.

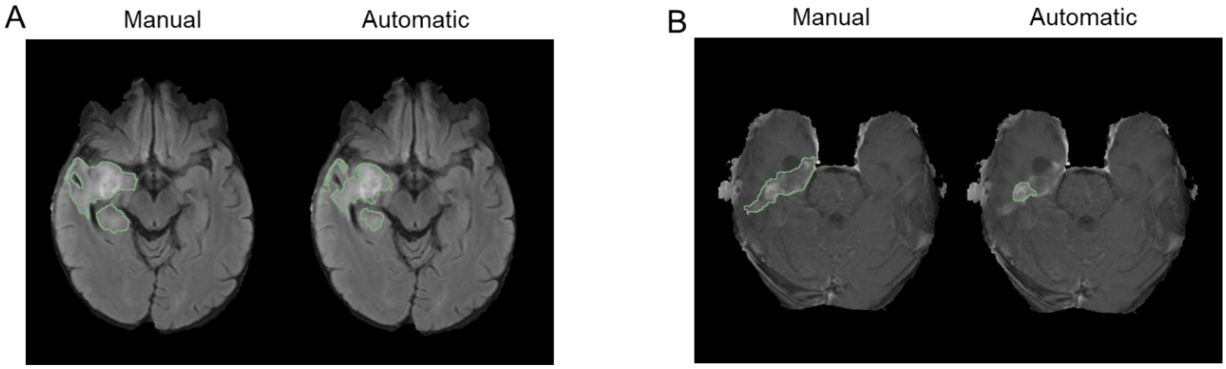


Figure 5-13. Examples of automatic segmentation with poor agreement with expert manual segmentation from the testing set is shown in (A) for FLAIR hyperintensity (overlaid on axial FLAIR image) and (B) contrast-enhancing (overlaid on T1 post-contrast image).

5.4.3 Repeatability of volume and RANO measurements in the post-operative patient cohort

Repeatability of manual and automatic measurements was assessed by comparing measurements from the two baseline visits for each patient. Comparing baseline visits 1 and 2 for FLAIR hyperintensity volume, the ICC was 0.983 ($p < 0.001$) for manual volume measurement and 0.986 ($p < 0.001$) for automatic volume measurement. For contrast-enhancing tumor volume, the ICC was 0.964 ($p < 0.001$) for manual volume measurement and 0.991 ($p < 0.001$) for automatic volume measurement.

Comparing baseline visits 1 and 2 for RANO measurements, the ICC was 0.984 ($p < 0.001$) for manual RANO and 0.977 ($p < 0.001$, Fig. 5-14) for AutoRANO. Notably, there were five patients assessed by one rater that had measurable lesions on one but not the other baseline visit. Similarly, there were three patients assessed by the other rater that had measurable lesions on one but not the other baseline visit. By comparison, when using the AutoRANO algorithm, no patients had a discrepancy in the presence/absence of measurable lesions between the two baseline visits.

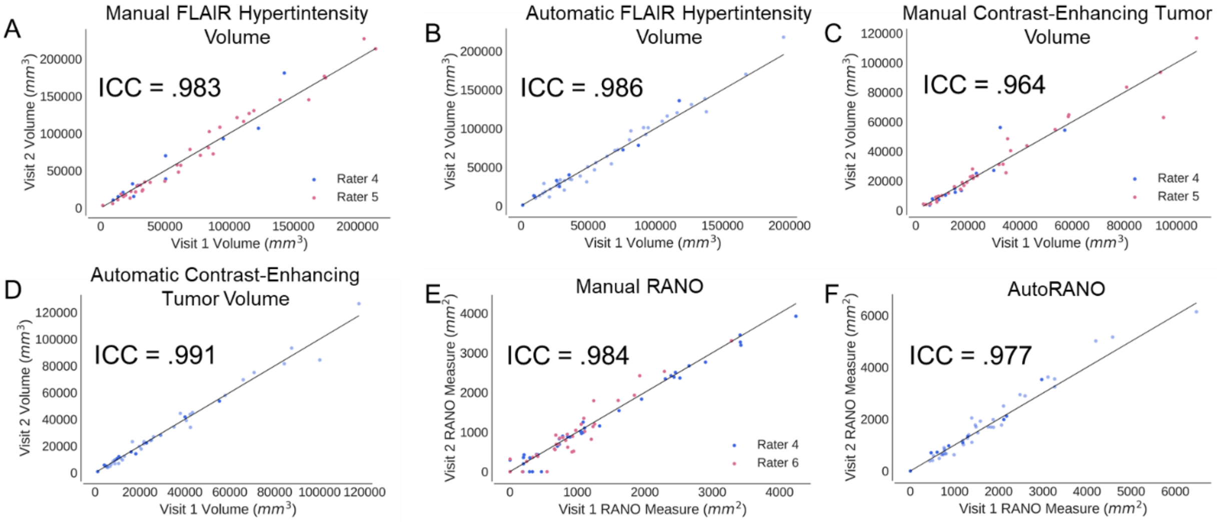


Figure 5-14. Volume and RANO measures are highly repeatable. Repeatability of (A) Manual FLAIR hypertintensity Volume, (B) Automatic FLAIR hypertintensity Volume, (C) Manual Contrast-Enhancing Tumor Volume, (D) Automatic Contrast-Enhancing Tumor Volume, (E) Manual RANO, and (F) AutoRANO in the post-operative patient cohort. Training and Testing Sets are shown as in light blue and dark blue, respectively, in B, D, and F. Line of identity (x = y) is shown in all plots.

5.4.4 Inter-rater agreement for manual RANO and agreement between manual RANO and AutoRANO

In assessing inter-rater agreement, the ICC for manual RANO measurements between the two expert raters was 0.704 ($p < 0.001$). In assessing rater-algorithm agreement, the ICC was 0.768 ($p < .001$) between AutoRANO and Rater 4 and 0.501 ($p < 0.001$, Fig. 5-15) AutoRANO and Rater 6.

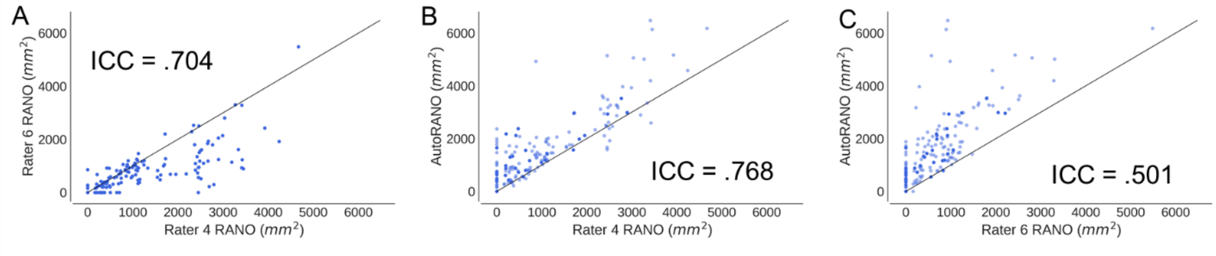


Figure 5-15. There was moderate inter-rater and manual-algorithm agreement for RANO measures. Agreement between RANO measures for (A) Rater 6 vs Rater 4, (B) AutoRANO vs Rater 4, and (C) AutoRANO vs Rater 6 in the post-operative patient cohort. Training and Testing Sets light blue and dark blue, respectively, in B and C. Line of identity (x = y) is shown in all plots.

5.4.5 Automatic treatment response assessment

Comparisons between nadir and the last patient visit were made (delta measure = last patient visit measure – nadir measure). In assessing rater-algorithm agreement for the delta measures, the ICC between automatic and manual delta measurements were 0.917 ($p < 0.001$), 0.966 ($p < 0.001$), and 0.850 ($p < 0.001$) for FLAIR hyperintensity volume, contrast-enhancing tumor volume, and RANO measures, respectively (Fig. 5-16). For non-progression vs progression, there was moderate agreement between both raters as well as Auto RANO: the κ coefficient for Rater 4 vs Rater 6, AutoRANO and Rater 4, and Auto RANO and Rater 6 was 0.546, 0.451, and 0.530, respectively (Fig. 5-17).

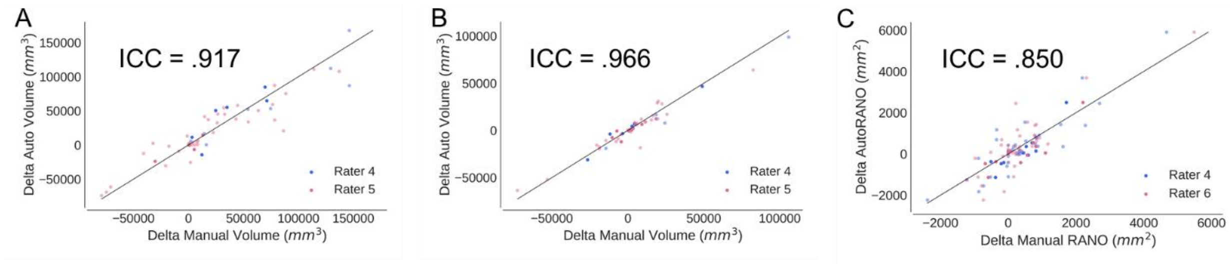


Figure 5-16. There was high agreement between manually and automatically derived longitudinal changes in volume and RANO measures. Agreement between automatic and manual delta measures for (A) FLAIR hypertintensity volume, (B) contrast-enhancing tumor volumes, and (C) RANO measure in the post-operative patient cohort. Training and Testing Sets are shown light blue/red and dark blue/red, respectively. Line of identity (x = y) is shown in all plots.

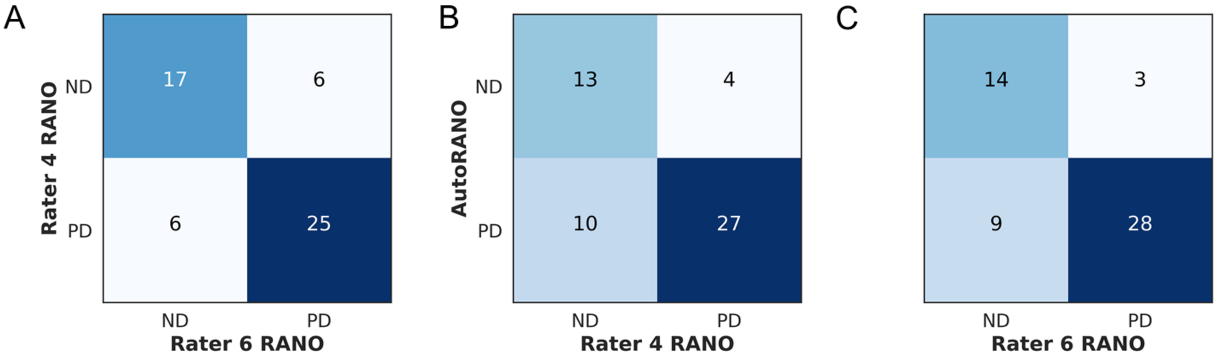


Figure 5-17. Classification of Non-Progressive Disease (ND) and Progressive Disease (PD) based on comparison of baseline and last patient visits for (A) Rater 4 vs Rater 6, (B) AutoRANO vs Rater 4, and (C) AutoRANO vs Rater 6 in the post-operative patient cohort.

5.4.6 Correlation between RANO measures and manual volume

Spearman’s ρ coefficient between manual RANO measures and manual enhancing-tumor volume was 0.787 ($p < 0.001$). Spearman’s ρ coefficient between AutoRANO measures and manual enhancing-tumor volume was 0.940 ($p < 0.001$, Fig. 5-18). Spearman’s ρ coefficient between delta manual RANO measures and delta manual enhancing-tumor volume was 0.744 ($p < 0.001$). Spearman’s ρ coefficient between delta AutoRANO measures and delta manual enhancing-tumor volume was 0.832 ($p < 0.001$, Supplementary Fig. 19).

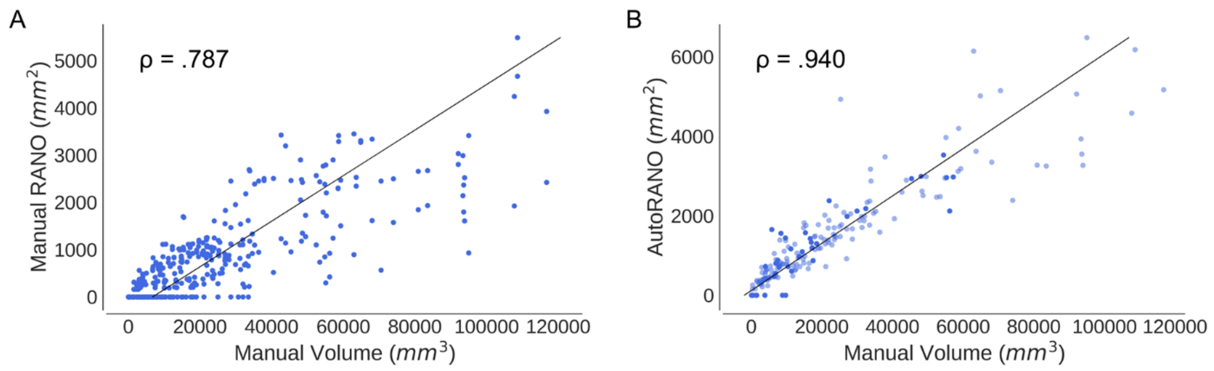


Figure 5-18. AutoRANO had higher agreement with manual contrast enhancing volume than manual RANO measures. Correlation between manual contrast-enhancing volume and RANO measures for (A) Manual RANO and (B) AutoRANO in the post-operative patient cohort. Training and Testing Sets are shown in light blue and dark blue, respectively in B. Linear fit is shown in all plots.

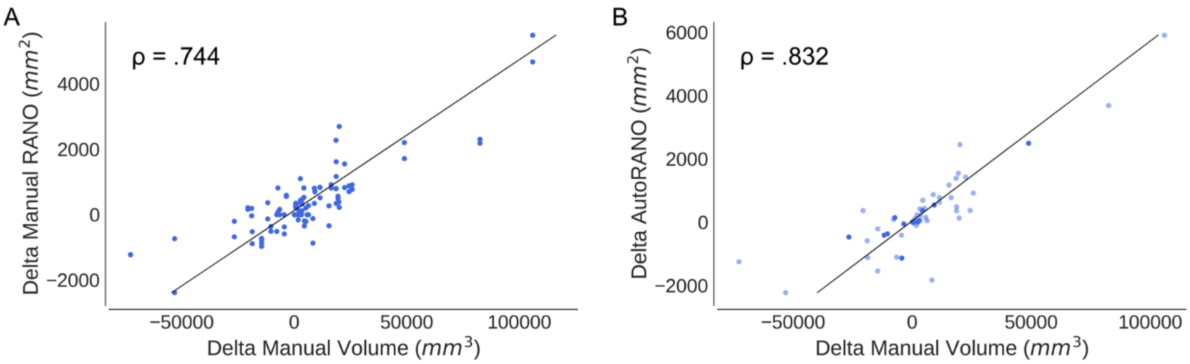


Figure 5-19. Delta AutoRANO had higher agreement with delta manual contrast enhancing volume than delta manual RANO measures. Correlation between delta manual contrast-enhancing volume and delta RANO measures for (A) Delta Manual RANO and (B) Delta AutoRANO in the post-operative patient cohort. Training and Testing Sets are shown in light blue and dark blue, respectively in B. Linear fit is shown in all plots.

5.5 Discussion

In this section, we demonstrate the utility of a fully automated, deep learning-based pipeline for calculation of tumor volumes and RANO measurements. A key image pre-processing step is brain extraction, which removes non-brain tissue – a significant source of error for downstream tumor segmentation. Although automatic brain extraction methods exist, performance can be compromised in a dataset with varying MR scanners and acquisition protocols, which in turn lead to variations in image contrast and intensity. Lesion pathology introduces an additional mode of variability due to the diversity of anatomical locations and radiographic features. Surgery also introduces variability due to the presence of blood products, proteinaceous material, resection cavities, and surgical defects/scars (Fig. 5-9). The result is that previous state-of-the art methods often require parameter tuning or manual correction on a case-by-case basis for patients with pathology, introducing time-consuming manual steps to the image pre-processing pipeline. Korfiatis et al. reported poor skull-stripping using an atlas-based method as a significant source of error for their automated tumor segmentation algorithm.[247] Here, we applied deep learning for brain extraction in a multi-institutional pre-operative glioma patient cohort with a wide variety of acquisition settings as well as in a post-operative glioblastoma patient cohort from a single institution. The U-Net outperformed other commonly used skull-stripping methods. Notably, no case-specific parameter tuning or editing was required to achieve high performance with the neural network. Our method outperformed Robust Learning-Based Brain Extraction, which is the only other brain extraction method studied that does not require any parameter tuning. Furthermore, the U-Net has robust recognition of tumor pathology, with a high mean fraction of tumor retained in the brain-extracted image.

After brain extraction, a deep learning framework was applied for FLAIR hyperintensity and contrast-enhancing tumor volume segmentation. Even with the varied acquisition protocols, our automatic pipeline proved to be robust for segmentation in the majority of patients in our multi-institutional dataset. We further developed an algorithm for automatic calculation of RANO measurements from contrast-enhancing tumor segmentations. In addition to the pre-operative setting, our algorithm demonstrated good performance in post-operative MRIs, which are particularly challenging given the frequent presence of surgical cavities and brain distortion. Furthermore, the algorithm was successfully applied in a longitudinal patient cohort including patients that had been treated with cediranib, which blunts the contrast enhancement, yielding ill-defined contrast enhancement margins that are difficult to contour. It is in these cases, particularly, that standardized segmentation is likely to be most helpful.

Based on the double baseline MRIs, both manually and automatically derived FLAIR hyperintensity volume, contrast-enhancing tumor volume, and RANO measurements were highly repeatable, showing intra-rater consistency. However, there were differences in inter-rater consistency. The RANO measurements from the AutoRANO algorithm were, on average, larger than those of the two human raters. This is likely due to the fact that our AutoRANO algorithm performs an exhaustive search of the longest perpendicular diameters while a human performs this estimation by eye, which is a less accurate method. This inaccuracy is further evidenced by the fact that the average RANO measurements differed between the two raters. In fact, consistent with prior reports on the variability in 2D measurements[238], it is not surprising that there was substantial variability between RANO measurements between our raters. In contrast, we found high agreement between manual raters and automatic volume for both contrast-enhancing tumor and FLAIR hyperintensity. This suggests that volume measurements allow for greater consistency across raters than RANO measurements.

There was high agreement between manual and automatic measures with regard to changes in tumor burden (both contrast-enhancing and FLAIR hyperintensity) during the course of longitudinal therapy. However, there was better agreement between manual raters and automated measurements for contrast-enhancing tumor volume compared to RANO measures. Thus, automated volume measurements were superior to AutoRANO measurements due to higher concordance with manual methods.

Interestingly, AutoRANO correlated better with manual contrast-enhancing tumor volume than the manual RANO measurements. Delta AutoRANO (the difference in the bi-dimensional measurements between the last visit and the nadir scan) also correlated better with delta manual contrast-enhancing tumor volume than delta manual RANO measurements. This suggests that AutoRANO may be a more accurate measure of tumor burden than manual RANO measurement in addition to the advantage of being fully automated.

One point to note is that the ICC values for manual vs automatic volumes were higher than the Dice scores for manual vs automatic segmentation. This is because Dice is a measure of the spatial overlap between the ground truth and segmentations, while the ICC compares volumes without considering spatial location. Both metrics provide useful but complementary information. Dice as a measure is more sensitive to differences in segmentation along the boundary of the lesion. Thus, if manual and automatic segmentations differed along the boundary, this can compromise the dice measure which is dependent on the degree of overlap. Furthermore, Dice coefficient can be sensitive to lesion size in that a few voxel difference in the location of the boundary can substantially reduce the Dice for small lesions but not as much for large lesions. In contrast, ICC of volume is less sensitive to boundary effects. If automatic segmentation was more conservative at some boundaries and more liberal at other boundaries compared to manual segmentation, these effects would cancel out and there would still be high concordance between manual and automatic volumes. Indeed, this is the case which is why the ICC values were higher than the Dice scores.

In the manual and automatic determination of response, there was moderate agreement between both raters as well as Auto RANO. There are two key reasons for this: 1) Because the cutoffs for each response category are percentage based, any variation in the average size of RANO measurements can dramatically affect categorization. There were indeed differences in the average size of the RANO measurements from Rater 1, Rater 2, and Auto RANO which affected response categorization. 2) Deciding if a small lesion around the 1 cm threshold is measurable is subjective. If a lesion is not measurable at nadir and becomes measurable on the last visit, this would be classified as progressive disease even if the change was only 1-2 mm. There were differences in the percentage of lesions considered measurable for both raters and

Auto RANO – even on the double baseline scans where size should not have changed. For these reasons, there was moderate rater to algorithm agreement in the call for progression, consistent with the moderate inter-rater agreement in the call of progression.

5.6 Limitations

There are some limitations to our study. First, the expert manual volume segmentations for each patient were derived from a single rater, which limits our ability to assess inter-rater variability of volume segmentation. Future studies could incorporate segmentations from multiple raters for segmentation. Second, our post-operative patient cohort contained imaging from only 54 patients from a single institution. Additional studies could utilize a larger, multi-institutional cohort and also assess performance early after surgery versus later after surgery as well as in responsive versus progressive disease. Third, our approach utilized a single neural network architecture without comparison with other approaches. Future work could explore the clinical utility of other neural network architectures as well as ensembles of neural network models to.[261] Furthermore, only patients with residual enhancing tumor of a certain size after surgery were enrolled in the clinical trials, which limits applicability to smaller tumors which may be harder to segment. Additionally, patient cohorts with 2D or 3D MR imaging was used in this study, as 3D MR imaging is not always available at all institutions. The utilization of only 3D MR imaging would further improve the reliability of bi-directional and volume measures.[241] Lastly, the confidence of the algorithm in its segmentations could be added to our pipeline to flag segmentations that require further verification from clinicians.[262,263] This would allow for more reliable integration into clinical workflows. Overall, our study shows that automated measures of tumor burden are highly reproducible and can reflect changes in tumor burden during the course of treatment. These automated tools could potentially be integrated in routine clinical care and imaging analyses performed as part of clinical trials and significantly enhance our accuracy in assessing treatment response.

5.7 Background on *IDH* mutations in glioma

In 2008, the presence of *IDH1*mutations, specifically involving the amino acid arginine at position 132, was demonstrated in in 12% of glioblastomas[100], with subsequent reports observing *IDH1* mutations in 50-80% of LGGs.[101] In the wild-type form, the *IDH* gene product converts isocitrate into α-ketoglutarate.[102] When *IDH* is mutated, the conversion of isocitrate is instead driven to 2-hydroxyglutarate, which inhibits downstream histone demethylases.[103] The presence of an *IDH* mutation carries important diagnostic and prognostic value. Gliomas with the *IDH1* mutation (or its homolog *IDH2*) carry a significantly increased overall survival than *IDH1/2* wild-type tumors, independent of histological grade.[100,104–106] Conversely, most lower grade gliomas with wild type IDH were molecularly and clinically similar to glioblastoma with equally dismal survival outcomes.[83] *IDH* wild-type grade III gliomas may in fact exhibit a worse prognosis than *IDH* mutant grade IV gliomas.[104] Its critical role in determining prognosis was emphasized with the inclusion of *IDH* mutation status as a classification parameter used in the 2016 update of WHO diagnostic criteria for gliomas.[107]

Pre-treatment identification of isocitrate dehydrogenase (*IDH)* status can help guide clinical decision making. First, a priori knowledge of *IDH1* status with radiographic suspicion of a low-grade glioma may favor early intervention as opposed to observation as a management option. Second, *IDH* mutant gliomas are driven by specific epigenetic alterations, making them susceptible to therapeutic interventions (such as temozolomide) that are less effective against

IDH wild-type tumors.[108,109] This is supported by *in vitro* experiments, which have found *IDH*-mutated cancer cells to have increased radio- and chemo-sensitivity.[110–112] Lastly, resection of non-enhancing tumor volume, beyond gross total removal of the enhancing tumor volume, was associated with a survival benefit in *IDH1* mutant grade III-IV gliomas but not in *IDH1* wild-type high-grade gliomas.[113] Thus, early determination of *IDH* status may guide surgical treatment plans, peri-operative counseling, and the choice of adjuvant management plans.

Non-invasive prediction of IDH status in gliomas is a challenging problem. A recent study by Patel *et al*. using MR scans from the TCGA/TCIA low-grade glioma database demonstrated that T2-FLAIR mismatch was a highly specific imaging biomarker for the *IDH*-mutant, 1p19q non-deleted molecular subtype of gliomas.[264] Other previous approaches toward prediction utilized isolated advanced MR imaging sequences, such as relative cerebral blood volume, sodium, spectroscopy, blood oxygen level-dependent, and perfusion.[265–270] An alternative radiomics approach has also been applied, which extracts radiographic features from conventional MRI such as growth patterns as well as tumor margin and signal intensity characteristics.[242,271] Radiomic approaches rely on multi-step pipelines that include generation of numerous pre-engineered features, selection of features, and application of traditional machine learning techniques.[272] Deep learning simplifies this pipeline by learning predictive features directly from the image. Deep learning has shown promising capabilities in prediction of key molecular markers in gliomas such as 1p19q codeletion and MGMT promoter methylation.[273–275] We hypothesize that a deep learning algorithm, using differences in radiographic differences between wild-type and mutants on conventional MR, can achieve high accuracy in predicting IDH mutation in gliomas (Fig. 5-20). In this section, we trained a deep learning algorithm to non-invasively predict IDH status within a multi-institutional dataset of low and high-grade gliomas.

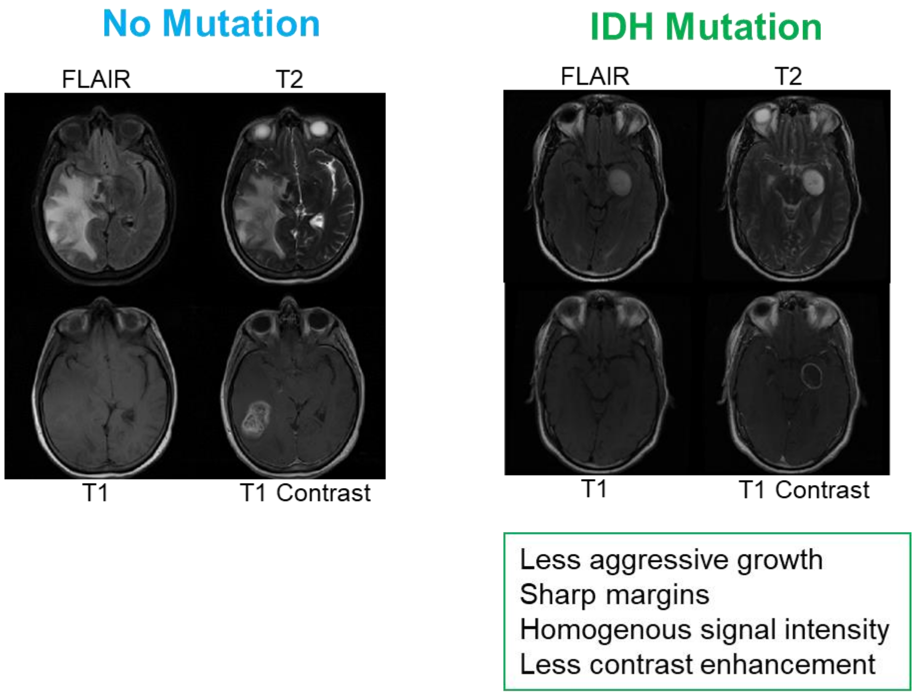


Figure 5-20. Radiographic differences between *IDH* wild-type and mutant on conventional MR

5.8 Materials and Methods

5.8.1 Patient Cohorts

We retrospectively identified patients with histologically confirmed World Health Organization grade II-IV gliomas with proven IDH status (after resection or biopsy) at the Hospital of the University of Pennsylvania (HUP), the Brigham and Women’s Hospital (BWH), and The Cancer Imaging Archive (TCIA). The study was conducted following approval by the HUP and DanaFarber/Brigham and Women's Cancer Center (DF/BWCC) Institutional Review Boards. MR imaging, clinical variables including patient demographics (i.e. age and sex), and genotyping data were obtained from the medical record under a consented research protocol approved by the DF/BWCC IRB. For the TCIA cohort, we identified glioma patients with preoperative MR imaging data from TCGA and IvyGap.[254] Under TCGA/TCIA data-use agreements, analysis of this cohort was exempt from IRB approval. All patients identified met the following criteria: (i) histopathologically confirmed primary grade II-IV glioma according to current WHO criteria, (ii) known IDH genotype, and (iii) available preoperative MR imaging consisting of pre-contrast axial T1-weighted (T1 pre-contrast), post-contrast axial T1-weighted (T1 post-contrast), axial T2-weighted fast spin echo (T2), and T2-weighted fluid attenuation inversion recovery (FLAIR) images. The scan characteristics for the 3 patient cohorts are shown in Fig. 5-21-23. Patients whose genetic data were not confirmed per criteria (see "Tissue Diagnosis and Genotyping" section below) were excluded. Our final patient cohort included 201 patients from HUP, 157 patients from BWH, and 138 patients from TCIA.

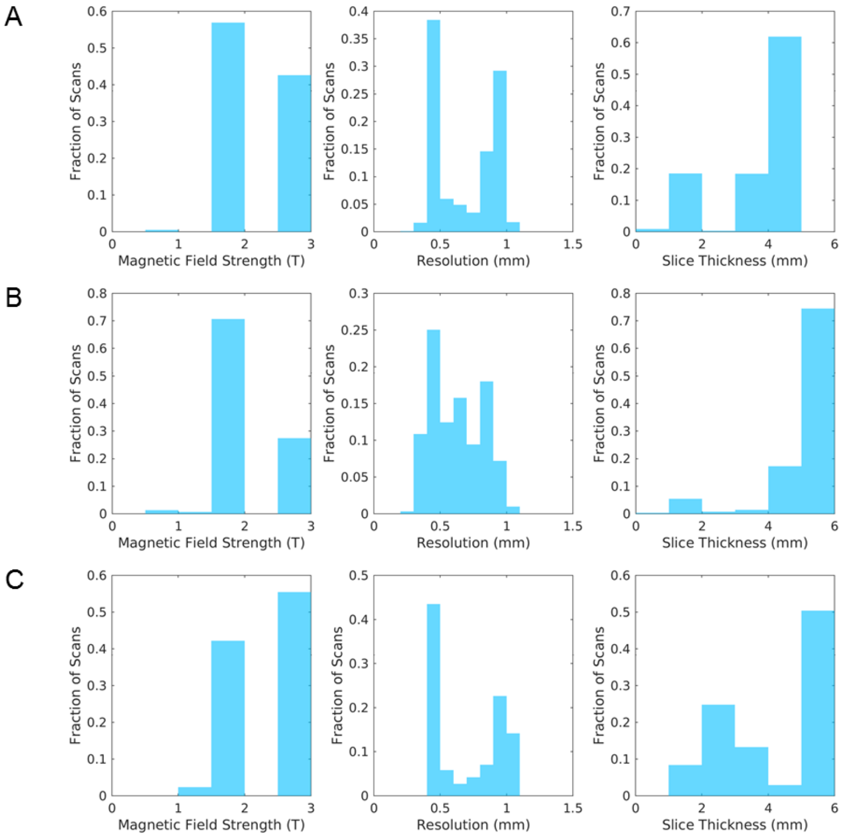


Figure 5-21. Magnetic field strength, resolution, and slice thickness of (a) HUP, (b) BWH, and (c) TCIA cohorts for FLAIR, T2, T1 pre-contrast, and T1 post-contrast MR images.

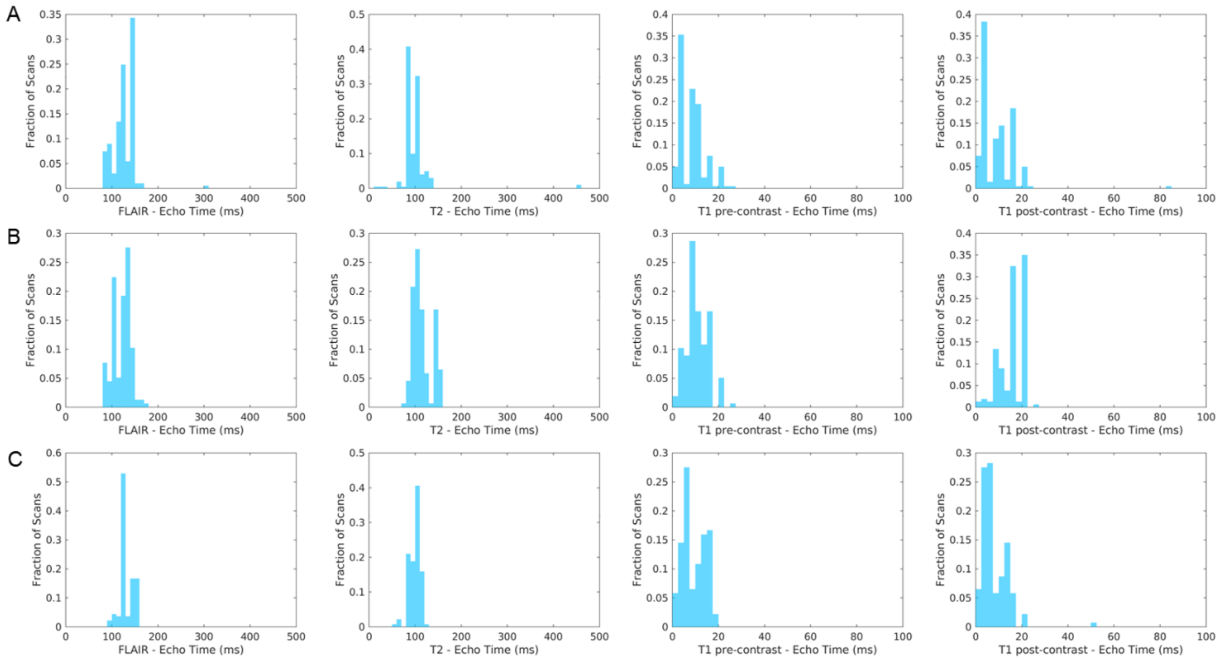


Figure 5-22. Echo time for FLAIR, T2, T1 pre-contrast, and T1 post-contrast MRI images of (a) HUP, (b) BWH, and (c) TCIA cohorts.

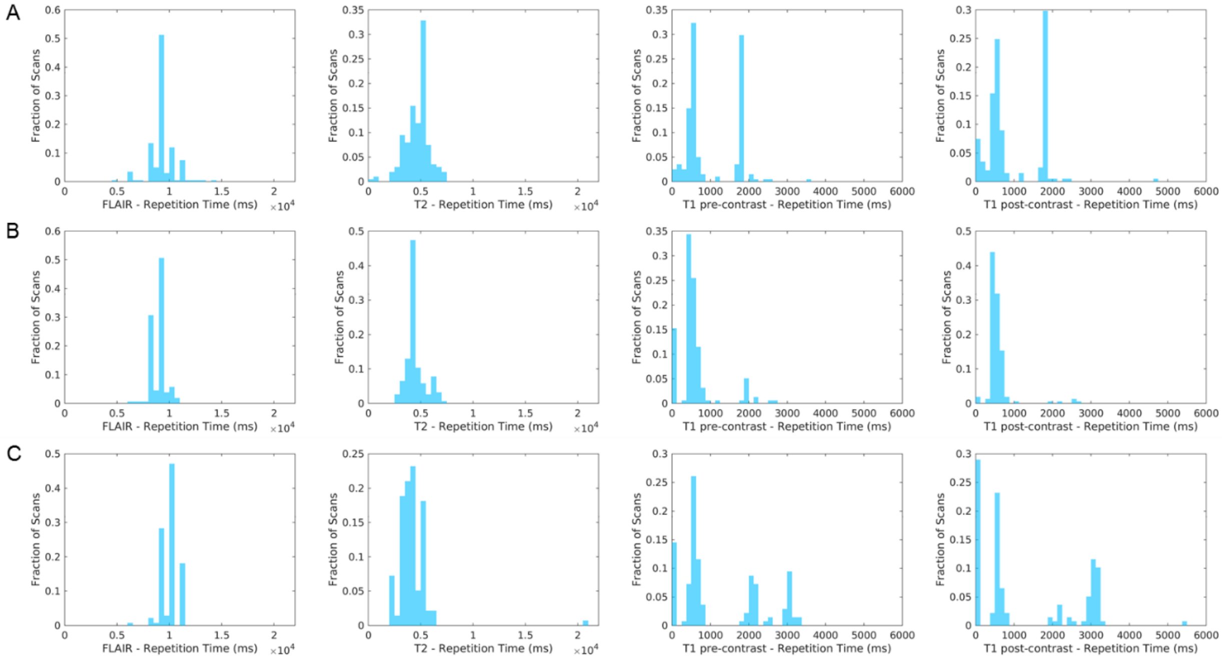


Figure 5-23. Repetition time for FLAIR, T2, T1 pre-contrast, and T1 post-contrast MRI images. of (a) HUP, (b) BWH, and (c) TCIA cohorts.

5.8.2 Tissue diagnosis and genotyping

For the HUP cohort, IDH1^{R132H} mutant status was determined using either immunohistochemistry (n = 93) or next-generation sequencing, performed by the Center for Personalized Diagnostics at HUP on 108 tumors diagnosed after February 2013. For the BWH cohort, *IDH*1/2 mutations were determined using immunohistochemistry, mass spectrometry-based mutation genotyping (OncoMap) [276], or capture-based sequencing (OncoPanel) [277,278] depending on the available genotyping technology at the time of diagnosis. OncoMap was performed by Center for Advanced Molecular Diagnostics of the BWH and Oncopanel was performed by Center for Cancer Genome Discovery of the Dana-Farber Cancer Institute. For patients under the age of 50 in the HUP and BWH cohorts, only gliomas with the absence of *IDH*1/2 mutation as determined by full sequencing assay were included in our analyses as *IDH* wild-type as to minimize the possibility of false negatives. *IDH*-mutated gliomas were defined by the presence of mutation as indicated by immunohistochemistry or sequencing on samples provided to the pathology department at each institution at the time of surgery. *IDH1*- and *IDH2*-mutated gliomas were collapsed into one category. For patients in the TCIA cohort, *IDH1*/2 mutation data were downloaded from TCGA and IvyGap data portal [254].

5.8.3 Tumor segmentation

For the HUP and TCIA cohorts, MR imaging for each patient was loaded into Matrix User v2.2 (University of Wisconsin, WI), and 3D regions-of-interest were manually drawn slice-by-slice in the axial plane for the FLAIR image by a user (H.Z.) followed by manual editing by a neuroradiologist (Q.S.). For the BWH cohort, tumor outlines were drawn with a user-driven, manual active contour segmentation method with 3D Slicer software (v4.6) on the FLAIR image (K.C.) and edited by an expert neuroradiologist (R.Y.H.) [258,279]. The segmented contour was then overlaid with source FLAIR, T2, T1 pre-contrast, and T1 post-contrast images.

5.8.4 Image pre-processing

All MR images were isotropically resampled to 1 mm with bicubic interpolation. T1 pre-contrast, T2, and FLAIR images were then registered to T1 post-contrast using the similarity metric. Resampling and registration was performed using MATLAB 2017a (Mathworks, MA). N4 bias correction (Nipype Python package) was applied to remove any low frequency intensity non-uniformity [257,280]. Skull-stripping was then applied from the FSL library to isolate regions of brain [248]. Image intensities were normalized by subtracting the median intensity of normal brain (non-tumor regions) and then dividing by the interquartile intensity of normal brain. To utilize information from all 3 spatial dimensions, we extracted coronal, sagittal, and axial tumor slices from each patient. Only slices with tumor were extracted. To extract a slice, a bounding rectangle derived from the tumor segmentation was drawn around the tumor. This ensures that the entire tumor area is captured as well as a portion of the tumor margin. Because every tumor is different in size, all slices were resized to 142x142 voxels for input into our neural network.

Gliomas are heterogeneous 3D volumes with complex imaging characteristics across each dimension. In our experiments, we choose to model this 3D heterogeneity by using 3 representative orthogonal slices, one each in the axial, coronal and sagittal planes. Together, these 3 orthogonal slices represent a single "sample" of the 3D tumor volume, and a total of three such samples were chosen for each patient based on the following scheme: 1) the coronal slice with the largest tumor area, the sagittal slice with the 75th percentile tumor area, and the axial slice with the 50th percentile tumor area, 2) the coronal slice with the 50th percentile tumor area, the sagittal slice with the largest tumor area, and the axial slice with the 75th percentile tumor area, 3) the coronal slice with the 75th percentile tumor area, the sagittal slice with the 50th percentile tumor area, and the axial slice with the largest tumor area. While each such sample may be somewhat correlated to other samples of the same tumor, gliomas exhibit marked heterogeneity and each additional set of orthogonal slices captures a marginal but significant amount extra information about that particular tumor. After pre-processing, the total number of patient samples was 603 for HUP, 414 for TCIA, and 471 for BWH. Image samples from the same patient were kept together when randomizing into training, validation, and testing sets. Another method of addressing overfitting is to augment the training data by introducing random rotations, translations, shearing, zooming, and flipping (horizontal and vertical), generating "new" training data [274]. The augmentation technique allows us to further increase the size of our training set. For every epoch, we augmented the training data before inputting it into the neural network. Augmentation was only performed on the training set and not the validation or testing sets. Data augmentation was performed in real time in order to minimize memory usage.

5.8.5 Residual neural network

Convolutional neural networks are a type of neural network developed specifically to learn hierarchical representations of imaging data. The input image is transformed through a series of chained convolutional layers that result in an output vector of class probabilities. It is the stacking of multiple convolutional layers with non-linear activation functions that allow a network to learn complex features. Residual neural networks won the 2015 Large Scale Visual Recognition Challenge by allowing effective training of substantially deeper networks than those used previously while maintaining fast convergence times [137]. This is accomplished via shortcut, “residual” connections that do not increase the network’s computational complexity [137]. Our residual network was derived from a 34-layer residual network architecture (Fig. 5-24A) [137]. As with the original residual network architecture, batch normalization was used after every convolutional layer [164]. Batch normalization forces network activations to follow a unit Gaussian distribution after each update, preventing internal covariate shift and overfitting [164]. The first two layers of the original residual network architecture, which sub-sample the input images, were not used, as the size of our input (142x142) is smaller than that of the original residual net input (224x224).

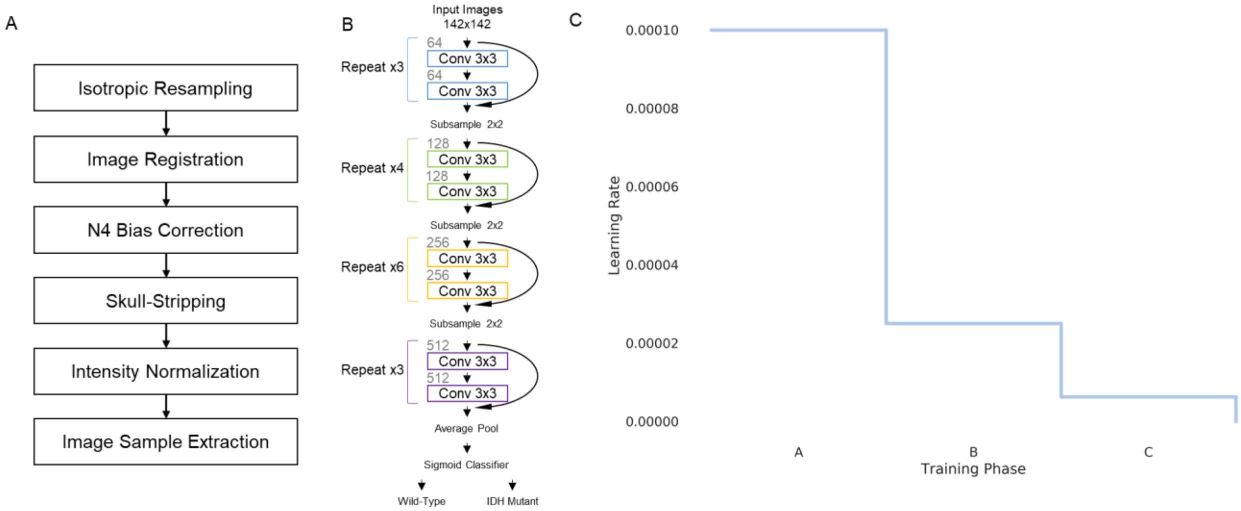


Figure 5-24. (A) Image pre-processing steps in our proposed approach. (B) A modified 34-layer residual neural network architecture was used to predict IDH status. (C) Displays the learning rate schedule. The learning rate was set to .0001 and stepped down to .25 of its value when there is no improvement in the validation loss for 20 consecutive epochs.

5.8.6 Implementation details

Our implementation was based on the Keras package with the TensorFlow library as the backend. During training, the probability of each patient sample belonging to the wild-type or mutant IDH class was computed with a sigmoid classifier. We used the rectified liner unit activation function in each layer. The weights of the network were optimized via a stochastic gradient descent algorithm with a mini-batch size of 16. The objective function used was binary cross-entropy. The learning rate was set to 0.0001 with a momentum coefficient of 0.9. The learning rate was decayed to 0.25 of its value after 20 consecutive epochs without an improvement of the validation loss. The learning rate was decayed 2 times (Training Phases A-C, Fig. 5-24B). At the end of training phase A and B, the model was reverted back to the model with the lowest validation loss up until that point in training. The final model was the one with the lowest validation loss at any point during training. Biases were initialized randomly using the Glorot uniform initializer [140]. We ran our code on a graphics processing unit to exploit its computational speed. Our algorithm was trained on a Tesla P100 graphics processing unit. Code for image pre-processing as well as trained models utilizing the modality networks heuristic can be found here: https://github.com/changken1/IDH_Prediction.

5.8.7 Training with three patient cohorts

Each patient cohort (HUP, BWH, and TCIA) was randomly divided into training, validation, and testing sets in an 8:1:1 ratio, balancing for mutation status and age. In our experiments training with all three patient cohorts, we combined HUP, BWH, and TCIA training sets. Similarly, we combined HUP, BWH, and TCIA validation sets as well as testing sets. The combined testing set was not disclosed until the model was finalized.

We implemented three different training heuristics. In the first heuristic, we input all sequences and dimensions into a single residual network with input size 12x142x142 (single combined network heuristic, Fig. 5-25A). In the second heuristic, we trained a separate residual network for each dimension (input size 4x142x142) and combined the sigmoid probabilities of each network with a logistic regression (dimensional networks heuristic, Fig. 5-25B). In the third heuristic, we trained a separate network for each MRI sequence (input size 3x142x142) and combined the sigmoid probabilities of each network with a logistic regression (sequence networks heuristic, Fig. 5-25C).

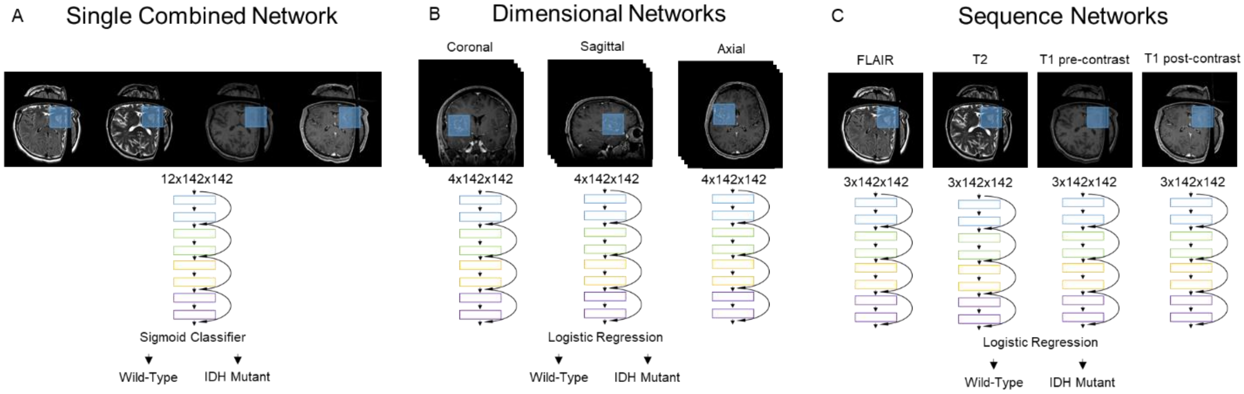


Figure 5-25. The training heuristics tested include a (A) single combined network, (B) dimensional networks, and (C) sequence networks. In the single combined network training heuristic, all sequences and dimensions were inputted into a single network. In the dimensional networks training heuristic, a separate network was trained for each dimension. In the sequence networks training heuristics, a separate network was trained for each MR sequence.

Because IDH status is correlated with age [242], we compared the results of residual neural networks with a logistic regression model based on age of patients in the training and validation sets. We also implemented a logistic regression model combining the sigmoid probability output of the residual neural networks and age.

5.8.8 Independent testing

We also trained residual networks with two patient cohorts with the goal of seeing if the model could predict *IDH* mutation status in the independent testing set without having been trained on any patients in that set. In these experiments, we combined the training sets of two patient cohorts. Similarly, we combined the validation sets and testing sets of two patient cohorts. The remaining patient cohort was kept aside as an independent testing set. The testing and independent testing sets were not disclosed until the final model was developed. The sequence networks training heuristic was used for these experiments.

5.8.9 Evaluation of models

The performance of models was evaluated by assessing the accuracy on training, validation, and testing sets. In addition, sigmoid or logistic regression probabilities were used to calculate Area Under Curve (AUC) of Receiver Operator Characteristic (ROC) analysis. Bootstrapping was used to calculate the confidence intervals (CI) of the AUC values.

5.9 Results

5.9.1 Patient characteristics

The median age of the HUP, BWH, and TCIA cohorts were 56, 47, and 52 years, respectively (Table 5-3). The percentage of males was 56%, 57%, and 57%, respectively. The HUP cohort was 19% grade II (72% *IDH*-mutant), 34% grade III (59% *IDH*-mutant), and 46% grade IV (3% *IDH* mutant). The BWH cohort was 20% grade II (100% *IDH*-mutant), 29% grade III (87% *IDH*-mutant), and 51% grade IV (26% *IDH* mutant). The TCIA cohort was 25% grade II (91% *IDH*-mutant), 32% grade III (70% *IDH*-mutant), and 43% grade IV (12% IDH mutant). Collectively, the HUP, BWH, and TCIA cohorts were 36%, 59%, and 50% *IDH*-mutant, respectively.

	HUP, n = 201	BWH, n= 157	TCIA, n= 138
Age	56 (18-88)	47 (18-85)	52 (21-84)
Sex (% Male)	56%	57%	57%
IDH mutation rate	36%	59%	50%
Grade & IDH status			
II Wild-Type	11	0	3
II Mutant	28	31	31
III Wild-Type	28	6	13
III Mutant	41	40	31
IV Wild-Type	90	59	53
IV Mutant	3	21	7

Table 5-3. Patient demographics, IDH status, and grade for HUP, BWH, and TCIA cohorts. Age is shown as median (minimum-maximum).

5.9.2 Optimization of deep learning model

We first determine the optimal training heuristics for the full multi-center data set by comparing three different heuristics (Fig. 5-26). A logistic regression model using age alone had an AUC of 0.88 on the Training set, 0.88 on the Validation set, and 0.89 on the Testing set (Table 5-4).

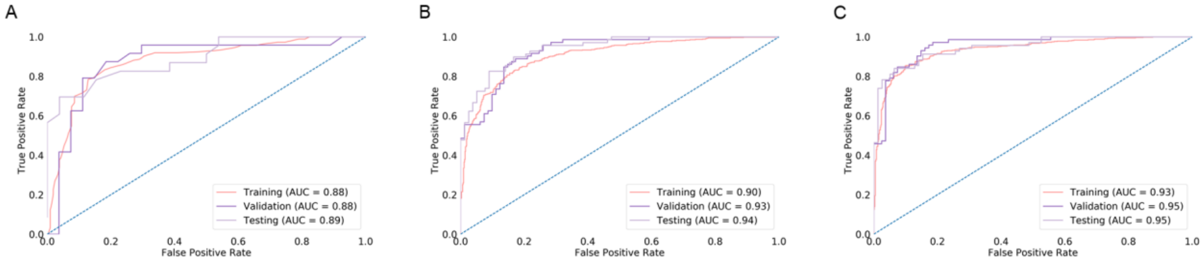


Figure 5-26. ROC curves for training, validation, and testing sets from training on three patient cohorts for (A) age only, (B) combining sequence networks, and (C) combining sequence networks + age. The testing set AUC for combing sequence networks + age was 0.95.

	Training Set HUP + BWH + TCIA n = 1188		Validation Set HUP + BWH + TCIA n = 153		Testing Set HUP + BWH + TCIA n = 147	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Age	82.6%	.88	82.4%	.88	79.6%	.89
Single Combined Network						
Single combined network	86.4%	.93	82.4%	.92	76.9%	.86
Single combined network + age	89.1%	.95	86.9%	.95	84.4%	.92
Dimensional Networks						
Coronal network	80.0%	.87	77.8%	.89	76.9%	.85
Sagittal network	78.8%	.86	79.1%	.88	79.6%	.86
Axial network	82.0%	.90	79.7%	.91	76.9%	.87
Combining dimensional networks	83.2%	.91	84.3%	.93	77.6%	.90
Combining dimensional networks + age	87.2%	.94	85.6%	.94	89.1%	.95
Sequence Networks						
FLAIR network	65.9%	.72	62.1%	.70	65.3%	.69
T2 network	68.4%	.74	66.0%	.77	67.3%	.73
T1 pre-contrast network	68.7%	.77	72.5%	.75	68.7%	.86
T1 post-contrast network	80.5%	.88	82.4%	.89	86.4%	.92
Combining sequence networks	82.8%	.90	83.0%	.93	85.7%	.94
T1C network + age	87.2%	.93	86.9%	.95	87.8%	.94
Combining sequence networks + age	87.3%	.93	87.6%	.95	89.1%	.95

Table 5-4. Accuracies and AUC from ROC analysis from training on three patient cohorts. The methods shown include age only, the single combined network training heuristic, the dimensional networks training heuristic, and the sequence networks training heuristic.

First, we constructed a single combined network model. After 157 epochs training, the resulting model had an AUC of 0.93 on the Training set, 0.92 on the Validation set, and 0.86 on the Testing set. When combined with age, the single combined network had improved performance with an AUC of 0.95 on the Training set, 0.95 on the Validation set, and 0.92 on the Testing set.

To demonstrate the individual predictive performance for different imaging dimensions, the coronal, sagittal, and axial networks were trained for 92, 82, and 122 epochs, respectively. The final model for the coronal, sagittal, and axial networks had Testing set AUCs of 0.85, 0.86, and 0.87, respectively. When the dimensional networks were combined, the AUC was 0.91 on the Training set, 0.93 on the Validation set, and 0.90 on the Testing set. Performance was improved when dimensional networks were combined with age with an AUC of 0.94 on the Training set, 0.94 on the Validation set, and 0.95 on the Testing set.

To demonstrate the individual predictive performance for different MRI sequences, the FLAIR, T2, T1 pre-contrast, and T1 post-contrast networks were trained for 88, 75, 76, and 325 epochs, respectively (Fig. 5-27). The final model for the FLAIR, T2, T1 pre-contrast, and T1 post-contrast networks had Testing set AUCs of 0.69, 0.73, 0.86, and 0.92, respectively. When the sequence networks were combined, the AUC was 0.90 on the Training set, 0.93 on the Validation set, and 0.94 on the Testing set. When sequence networks were combined with age the AUC was 0.93 on the Training set, 0.95 on the Validation set, and 0.95 on the Testing set (Fig. 5-26). Looking at predictive performance for the individual tumor grades, the AUC for the Validation and Testing cohorts were 0.85 (n = 66), 0.91 (n = 81), and .94 (n = 153) for grades 2, 3, and 4, respectively.

Overall, combining the sequence networks and age resulted in the highest performance in terms of accuracy and AUC values in the validation and testing set. This approach was subsequently applied to independent data set testing.

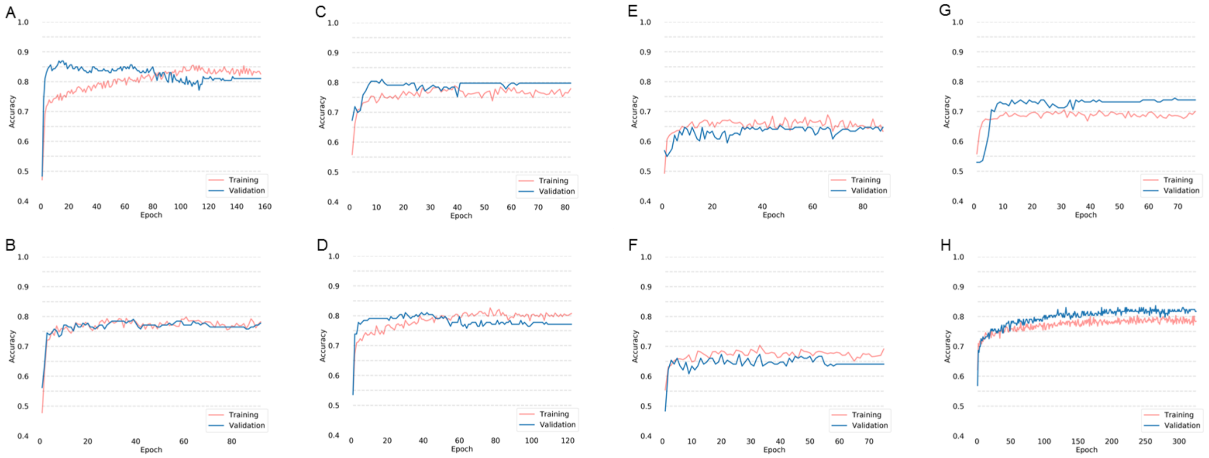


Figure 5-27. Training and validation accuracy for three patient cohort training for (A) single combined network, (B) coronal network, (C) sagittal network, (D) axial network, (E) FLAIR network, (F) T2 network, (G) T1 pre-contrast network, and (H) T1 post-contrast network. Training accuracy shown is for augmented training data.

5.9.3 Training on two patient cohorts and independent performance testing on the third cohort

To examine the generalizability of our model, the sequence network training heuristic was applied to training on two patient cohorts at a time. FLAIR, T2, T1 pre-contrast, and T1 post-contrast residual networks were trained on the combined Training sets of HUP + TCIA, HUP + BWH, and TCGA + BWH with data from the remaining site reserved for independent testing (Table 5-5). The average AUCs for combining sequence networks within the Training, Validation, Testing, and Independent Testing Cohorts were 0.90 (95% CI 0.88-0.92), 0.89 (95% CI 0.84-0.94), 0.92 (95% CI 0.88-0.96), and 0.85 (95% CI 0.82-0.88), respectively. When age was combined with sequence networks, the average AUCs were 0.94 (95% CI 0.92-0.95), 0.95 (95% CI 0.91-0.98), 0.95 (95% CI 0.91-0.98), and 0.91 (95% CI 0.88-0.93) respectively within the Training, Validation, Testing, and Independent Testing sets.

Comparatively, a logistic regression model utilizing age alone had an average AUC of 0.88, 0.88, 0.89, and 0.87 respectively within the Training, Validation, Testing, and Independent Testing sets. The average accuracy, sensitivity, and specificity for combined model for age and sequence networks on the independent Testing set was 82.1%, 79.1%, and 87.0%, respectively.

	Training Set		Validation Set		Testing Set		Independent Testing Set	
	HUP + TCIA, n = 813		HUP + TCIA, n = 102		HUP + TCIA, n = 102		BWH, n = 471	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Age	83.0%	.89	73.5%	.83	82.4%	.86	77.7%	.86
FLAIR network	68.3%	.74	63.7%	.68	70.6%	.75	58.2%	.70
T2 network	62.4%	.64	64.7%	.73	69.6%	.64	62.0%	.72
T1 pre-contrast network	72.8%	.81	76.5%	.76	77.5%	.88	58.8%	.76
T1 post-contrast network	83.5%	.90	83.3%	.93	87.3%	.96	65.0%	.82
Combining modality networks	83.8%	.91	80.4%	.92	90.2%	.97	67.1%	.86
Combining modality networks + age	89.4%	.95	79.4%	.92	91.2%	.97	77.5%	.90
	HUP + BWH, n = 858		HUP + BWH, n = 111		HUP + BWH, n = 105		TCIA, n = 414	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Age	83.6%	.88	83.8%	.91	82.9%	.92	78.3%	.85
FLAIR network	69.1%	.77	60.4%	.72	71.4%	.75	67.6%	.73
T2 network	71.2%	.77	65.8%	.78	66.7%	.74	55.8%	.56
T1 pre-contrast network	68.8%	.78	76.6%	.73	72.4%	.88	65.5%	.74
T1 post-contrast network	76.5%	.85	74.8%	.83	81.0%	.89	79.5%	.85
Combining modality networks	81.4%	.90	78.4%	.89	83.8%	.91	79.0%	.87
Combining modality networks + age	86.7%	.94	85.6%	.96	89.5%	.94	84.5%	.91
	TCIA + BWH, n = 705		TCIA + BWH, n = 93		TCIA + BWH, n = 87		HUP, n = 603	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Age	81.3%	.86	83.9%	.89	82.8%	.88	84.1%	.90
FLAIR network	67.2%	.75	66.7%	.69	63.2%	.64	63.5%	.65
T2 network	69.2%	.78	68.8%	.71	69.0%	.75	65.3%	.73
T1 pre-contrast network	69.4%	.79	68.8%	.73	63.2%	.83	72.5%	.75
T1 post-contrast network	75.0%	.82	77.4%	.85	81.6%	.89	68.5%	.85
Combining modality networks	80.4%	.89	77.4%	.87	77.0%	.89	74.8%	.83
Combining modality networks + age	85.4%	.92	87.1%	.96	89.7%	.94	84.1%	.91
	Average		Average		Average		Average	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Age	82.7%	.88	80.4%	.88	82.7%	.89	80.5%	.87
FLAIR network	68.3%	.75	63.4%	.70	68.7%	.72	63.0%	.69
T2 network	67.6%	.73	66.3%	.74	68.4%	.71	61.6%	.68
T1 pre-contrast network	70.3%	.79	74.2%	.74	71.4%	.87	66.2%	.75
T1 post-contrast network	78.5%	.86	78.4%	.87	83.4%	.91	70.4%	.84
Combining modality networks	81.9%	.90	78.8%	.89	84.0%	.92	73.5%	.85
Combining modality networks + age	87.2%	.94	84.0%	.95	90.1%	.95	82.1%	.91

Table 5-5. Training, validation, testing, and independent testing set performance from training on two patient cohorts using the sequence networks training heuristic.

5.10 Discussion

In this section, we demonstrate the utility of deep learning to predict *IDH* mutation status in a large, multi-institutional dataset of gliomas as part of a larger effort to apply deep learning techniques to the field of neuro-oncology. To our knowledge, this is the largest study to date on the prediction of *IDH* status from conventional MR imaging and deep learning methods. Furthermore, our algorithm has broad applicability by utilizing conventional MR performed at different institutions, as advanced MR sequences or other modalities may not be part of the standard imaging protocol. Pre-treatment identification of *IDH* status may be important in clinical-decision making as it may guide patient management, choice of chemotherapy, and surgical approach.

We did not include WHO grade information in our prediction model since this data would not have been known *a priori* without pathological tissue after invasive biopsy or surgery. The goal of our algorithm was to use conventional MR sequences to predict IDH mutation status before surgery. Furthermore, we did not train separate networks for each tumor grade to reflect the pre-operative clinical scenario, when the WHO grade remains unknown prior to acquisition of pathological tissue from biopsy or surgery. Increasing research and the updated 2016 WHO classification of CNS tumors further emphasize molecular phenotype as a critical determinant of glioma behavior even before the assignment of histopathologic grade [107].

Previous studies have reported an association between radiographic appearance and *IDH* genotype within gliomas. *IDH* wild-type grade II gliomas are more likely to display an infiltrative pattern on MRI, compared to the sharp tumor margins and homogenous signal intensity characteristic of *IDH* mutant gliomas [281]. Patel *et al*. found T2-FLAIR mismatch to be a specific biomarker for *IDH*-mutant, 1p19q non-deleted gliomas [264]. Hao et al. scored pre-operative MRIs of 165 patients from the TCIA/TCGA according to the Visually Accessible Rembrandt Images (VASARI) annotations and found that increased proportion of necrosis and decreased lesion size were the features most predictive of an IDH mutation [245]. However, VASARI features overall achieved lower accuracy than texture features in this study. In another study of 153 patients with glioblastoma using the VASARI features, Lasocki et al. found that if a particular glioblastoma does not have a frontal lobe epicenter and has less than 33% non-enhancing tumor, it can be predicted to be *IDH1*-wildtype with a high degree of confidence [282]. One significant limitation of this study is that only five glioblastoma patients had *IDH1* mutation (3.3%). Furthermore, Yamashita et al. found that mutant *IDH1* glioblastoma patients had a lower percentage of necrosis within enhancing tumor with the caveat that the study included only 11 *IDH1* mutant tumors [283].

As such, various studies have used a radiomics approach to predict *IDH* status. Zhang et al. used clinical and imaging features to predict *IDH* genotype in grade III and grade IV gliomas with an accuracy of 86% in the training cohort and 89% in the validation cohort [242]. Hao et al. used preoperative MRIs of 165 MRIs from the TCIA to predict *IDH* mutant status with an AUC value of 0.86 [245]. Similarly, Yu et al. used a radiomic approach to predict *IDH* mutations in grade II gliomas with an accuracy of 80% in the training cohort and 83% on the validation cohort [284]. Deep learning simplifies the multi-step pipeline utilized by radiomics by learning predictive features directly from the image, allowing for greater reproducibility. In this study, we demonstrate that accurate prediction can be achieved in a multi-institutional patient cohort of both low- and high-grade gliomas without pre-engineered features.

One challenge of training deep neural networks is the need for a large amount of training data. We addressed this by artificially augmenting our imaging data, in real-time, before each training epoch. This has the additional benefit of preventing overfitting, which is another common issue when training networks. We also utilized batch normalization after each convolutional layer to prevent overfitting, as with the original residual network architecture. We implemented various training heuristics with training on three patient cohorts – namely a single combined network, dimensional networks, and sequence networks. Under the dimensional networks training heuristic, we trained a neural network for coronal, sagittal, and axial dimensions which had similar testing set performance. These results suggest that all dimensions have similar predictive value. Under the sequence networks training heuristic, we trained a neural network for each MR sequence. Notably, T1 post-contrast images conferred a higher predictive value compared to other MR sequences and appeared to drive the vast majority of the accuracy of the combined sequence model with additional sequences contributing a smaller incremental benefit. Indeed, the contrast-enhancing regions were also highlighted when Grad-CAM visualization was used to highlight the salient features used in prediction of the single combined model (Fig. 5-28).[285] The only imaging-only models that outperformed the age-only logistic regression model in terms of accuracy in the validation and testing set were the T1 post-contrast network and a model combining sequence networks. Overall, a combination of sequence networks and age offered the highest accuracy in the validation and testing sets.

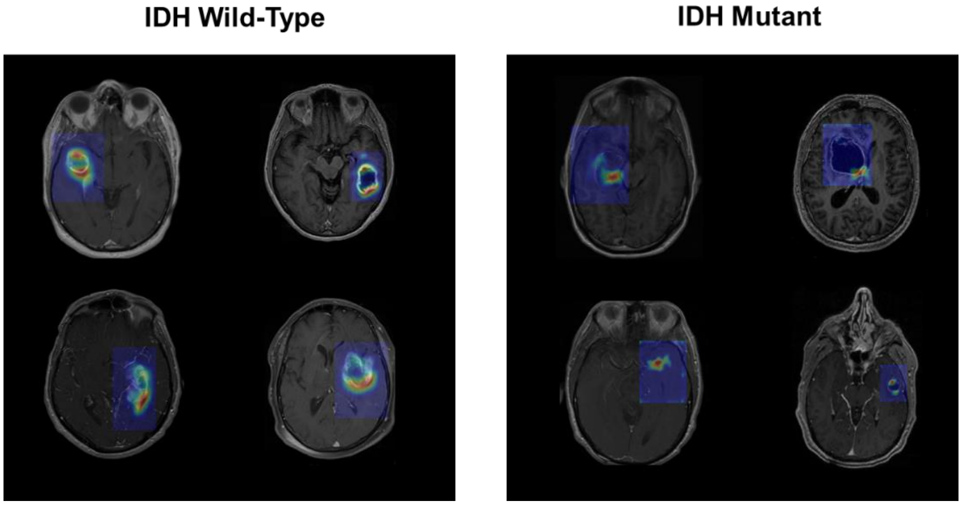


Figure 5-28. Grad-CAM visualizations of two patients in the independent test set, highlighting the salient features used by the neural network model

When the sequence networks training heuristic was applied to training on two patient cohorts at a time, similar results were observed when training on three patient cohorts. For training on HUP + TCIA, HUP + BWH, and TCIA + BWH, combining sequence networks and age had a higher AUC than a logistic regression using age only in the training, validation, testing, and independent testing sets. However, the AUC of the combined sequence network and age model within the independent testing set was lower than that of the testing set. The most likely reason for this are the differences in scan parameters and in *IDH* mutation rate among the different patient cohorts. In the ideal scenario, all patient scans would be collected with consistent acquisition parameters (field strength, resolution, slice thickness, echo time, and repetition time), and *IDH* mutation rate would be the same. However, this would be challenging in practice, as MR scanner models and acquisition parameters, as well as the demographics of patient captured, vary widely from institution to institution. Our approach distinguishes itself from past studies in the field by using multi-institutional data and makes an important first step towards achieving the goal of independent validation, which is necessary if radiogenomic tools are to be used in a clinical setting.

5.11 Limitations

There are several possible improvements to our approach. First, the potential of advanced MR sequences in the prediction of *IDH* genotype has been demonstrated in several studies [265–269]. We did not utilize such sequences, but future studies can combine advanced imaging modalities with conventional MR imaging to test for possible enhancement of prediction performance. However, addition of these advanced MR sequences is also a limitation in that these sequences may not be available at every institution. Second, sufficient cohort size is a limiting factor in the training of deep neural networks. Although we overcame this partially though data augmentation and extracting multiple imaging samples from the same patient, it is likely a larger patient population would further improve algorithm performance, especially given the heterogeneity in image acquisition parameters. Third, the use of other techniques such as dropout, L1/L2 regularization, and augmentation via generative adversarial networks may improve the generalizability of our model[182,286], although we found that basic data augmentation and batch normalization were sufficient to prevent overfitting of our model, as evidenced by the high testing accuracies. Lastly, incorporation of spatial characteristics of IDH-mutated gliomas (such as unilateral patterns of growth and localization within single lobes) into the deep neural network may further improve model performance.[281]

5.12 Future directions

Beyond the applications described in this chapter, there are many potential tools that would be helpful in assisting clinical decision making. In this section, we describe a non-invasive approach to prediction *IDH* mutation status. Other studies have show the capacity for non-invasive prediction of other molecular markers such as 1p19q codeletion and O6-methylguanine-DNA methyltransferase (*MGMT*) methylation.[287–289] Such molecular markers can be determined via pathological assessment of the tumor specimen that is acquired during the initial surgery and is not expected to change in remaining unresected tumor and during tumor recurrence.[290] However, the aforementioned molecular markers represent an incomplete picture of the genetic landscape of the tumor that determines response to chemotherapy. Rather, GBMs are more accurately described as “many tumors in one” with significant genetic heterogeneity within a given tumor focus.[291] This heterogeneity has been implicated as a driver of tumor progression, growth, and resistance to therapy.[292] Studies have shown that there are genetically distinct clonal populations with different driver mutations (such as in *CDKN2A, TP53, EGFR, PTEN, and NF1*) that may confer different therapeutic sensitivities.[100,293–295] This is further complicated by primary or recurrent tumors with multiple foci, which may have different genetic characteristics.[296] Additionally, temozolomide can induce adaptive genetic changes that confer resistance to treatment.[297,298] Thus, multi-focal and repeated tumor biopsies are needed to assess the genetic characteristics of the tumor. However, this is not feasible due to surgical risk, patient burden, and/or inoperable location. Future studies can use deep learning applied to multi-parametric Magnetic Resonance (MR) to non-invasively assess key driver mutations that vary spatially and temporally during the course of treatment.

Furthermore, we described in the section an approach for automated treatment response assessment for glioma. What is also needed is the ability to predict drug response before the treatment is even applied. An example would be bevacizumab, which has been shown to work for a subset of patients while ineffective for others.[299,300] Previous studies have shown the potential of a radiomics approach for prediction response to bevacizumab.[60,301] A natural extension would be to use a deep learning approach to predict drug response to bevacizumab and other therapies. The predictive capacity of these algorithms can be further augmented with the use of advanced MR modalities, which can elucidate vascular structure that can impact drug delivery.[302,303] Lastly, our approach focused our application on gliomas, but our algorithms can be extended to other diseases as well such as pediatric tumors and brain metastases.[304,305]

5.13 Conclusions

We developed an open-source, fully automatic pipeline for brain extraction, tumor segmentation, and RANO measurements and applied it to a large, multi-institutional pre-operative and post-operative glioma patient cohort. We showed that automated volume and AutoRANO measurements are highly reproducible and are in agreement with human experts in terms of change in tumor burden during the course of treatment. This tool may be helpful in clinical trials and clinical practice for expediting measurement of tumor burden in the evaluation of treatment response, decreasing clinician burden associated with manual tumor segmentation and decreasing inter-observer variability. Furthermore, our algorithm serves as a proof-of-concept for automated tools in the clinic and demonstrates their applicability to other tumor pathologies. We also developed a technique to non-invasively predict *IDH* genotype in grade II-IV glioma using conventional MR imaging. In contrast to a radiomics approach, our deep learning model does not require pre-engineered features. Our model may have the potential to serve as a noninvasive tool that complements invasive tissue sampling, guiding patient management at an earlier stage of disease and in follow-up.

6 Enhancing ischemic stroke workflows with computational tools

6.1 Potential for automated tools for ischemic stroke

The current patient workflow for the patient with symptoms suspicious of stroke is to receive either CT and/or MR imaging, depending on the protocol of the specific hospital. Based on the information revealed by physical examination and radiology, the patient receives a treatment that usually consists of either tPA (if within 4.5 hours of stroke onset) or mechanical thrombectomy. The patient is then monitored and treated for any additional clinical symptoms until the patient recovers (or deteriorates, in the unfortunate cases). Within this clinical workflow, there are many potential opportunities for automated tools. The ones that I will focus on in this chapter are: 1) Detection of stroke and volumetric assessment, 2) Prediction of patient outcomes, and 3) Risk assessment of associated clinical symptoms.

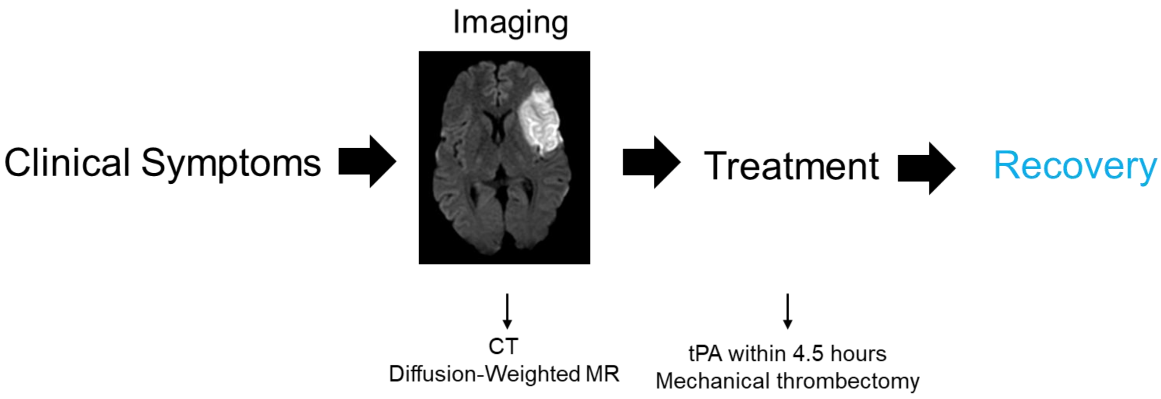


Figure 6-1. The current clinical workflow for patient with suspected ischemic stroke.

6.2 Background on automatic segmentation

The use of artificial intelligence in neurological disorders has significantly evolved over the past few years with introduction of several novel automated systems. Examples include AI systems used in neuro-critical care units to monitor hemodynamics[306] and deep-learning based algorithms for automatic seizure detection.[307] One active area of interest has been on differentiating structural brain lesions such as brain infarcts, hemorrhages, tumors, and MS plaques from the surrounding brain tissue.[8,42,308,309] Stroke represents an attractive model to develop a segmentation algorithm because it is a prevalent condition. This allows for testing and validating developed algorithms in large number of patients, which is critical for generalizability.

Accurate estimation of the infarct volume is invaluable because quantitative measurement of infarct volume rather than qualitative assessment of infarct size injects statistical power to stroke research. However, manual delineation of stroke regions is time-consuming, and subject to inter-rater variability.[310] Furthermore, segmentation is a highly difficult task as there can be variability in size and location of infarcts as well as ill-defined boundaries. As such, there has been efforts to develop automatic methods of performing infarct segmentation. Existing methods often require multiple imaging modalities (such as T1-weighted, T2-weighted, and fluid attenuation inversion recovery (FLAIR)), are semi-automatic and thus require manual input, or are constrained by cost and convenience.[118,119,311] Specifically, the need for multiple modalities can compromise the method if there is a defect in acquisition, such as the presence of an imaging artifact. Diffusion Weighted MR Imaging (DWI) is frequently used in patients with stroke to evaluate the extent of irreversibly damaged ischemic tissue.[312] Recently, Chen et al. developed a fully automatic method using convolutional neural networks using DWI with the limitation that their approach only utilizes information in only 2 dimensions, ignoring lesion information within the axial plane.[121] In this study, we developed a 3-dimensional deep learning approach for ischemic stroke volumetric segmentation utilizing only DWI imaging and a large clinical dataset of 1,205 consecutive patients. Furthermore, we demonstrated that both manual and automatically segmented volumes could be used to predict functional outcome and survival.[313] To facilitate the use of algorithms developed in this work (entitled Deep Learning Tool for Ischemic Stroke (DeLTIS)), we have made the code as well as trained model publically available for the larger research and clinical community to use.

6.3 Methods

6.3.1 Patient Cohort

The study was conducted following approval by the Partners Healthcare Human Studies Committee. Informed consent was obtained from each patient or from the patient’s relatives. Clinical and imaging data were collected prospectively in 1,205 consecutive patients from the inpatient population of the Massachusetts General Hospital (MGH) with DWI-confirmed acute ischemic stroke recruited between June 2009 and December 2011. Exclusion criteria included patients with intracerebral hemorrhage on the acute neuroimaging study, patients with contraindications to MRI, and patients who were admitted after 72 hours of symptom onset.

One investigator blinded to DWI findings performed outcome assessments through in person evaluations, phone interviews, or reviews of physician notes obtained during outpatient visits when the patient was unavailable for a follow-up visit. Follow-up evaluation included assessment of survival and functional outcome using modified Rankin Score dichotomized as good (mRS <=2) and bad (mRS >2) outcome at 90 ± 15 days. We used the Social Security Death Index to confirm the survival status at 90 days in patients who were not available for follow-up.

6.3.2 MR imaging

DWI (b-value = 0, b0, and 1000 s/mm², b1000) was performed on 1.5T General Electric (Milwaukee, WI) and 1.5T Siemens (Erlangen, Germany) MR instruments. The full diffusion tensor was sampled using a single-shot echo-planar imaging (EPI) technique repeated in at least six non-collinear diffusion gradient directions. The resolution in the x-y plane ranged from 0.859-1.72 mm and the resolution in the z plane ranged from 6-6.5 mm.

Manual annotations of acute infarcts were generated using an image outlining software (MRICron, United States) by one investigator (JH) who had access to clinical stroke characteristics of the patients and used this information to identify whether a given hyperintensity on DWI was clinically relevant. All outlines were adjudicated by a second investigator (HA) and final outlines were generated.

The patients were randomly divided into Training (n = 720), Validation (n=243), and Testing (n=242) sets in a 3:1:1 ratio. The Training Set was used to train our deep-learning algorithm and the performance of Validation Set was evaluated to assess for under/overfitting. Training was stopped when performance on the Validation Set no longer improved. The Testing Set was used for evaluation of segmentation quality once the model was finalized to ensure generalizability of the trained model.

6.3.3 Data pre-processing for segmentation

For pre-processing, we normalized the intensity of each b0 and b1000 DWI image to zero mean and unit variance. Notably, we did not resample the images in order to avoid introducing any resampling errors in both the image and manual segmentations. Furthermore, we did not apply brain extraction to the images, after observing that our neural networks can accurately segment the stroke lesions without this preprocessing step.

6.3.4 Neural Network Architecture for Segmentation

We utilized the 3D U-Net architecture, a network designed for fast and precise segmentation (Fig. 6-2A), implemented within the DeepNeuro framework.[10,259] A 3D approach was chosen instead of a 2D approach because a 3D approach incorporates information between adjacent axial slices. Because lesion segmentation is a challenging problem, modifications of the 3D U-Net architectures were investigated. Specifically, we individually incorporated residual connections, inception modules, dense connections, and squeeze-and-excitation modules (Fig. 6-2B) into the standard 3D U-Net architecture, all of which are state-of-the-art components that have improved neural network architectures for classification tasks. Residual connections are "shortcut" connections that allow for skipping of convolution layers.[137] Inception modules have multiple pathways with different convolution filter sizes, allowing the network to learn from a variety of field-of-views.[163] Dense connections allow feature maps from every convolutional layer to be carried into successive layers.[212] Squeeze-and-excitation modules allow relationships to be learned between different feature maps at different layers of the neural network by rescaling a layer with compressed feature maps from the previous layer.[314] Here, we utilized Inception modules from the Inception-V4 architecture.[163] Dense connections with a growth rate of two were used in place of each block of convolutions. Squeeze-and-excitation modules with a reduction ratio of 16 was utilized to reduce the computational complexity. The architectures consists of a downsampling and an upsampling arm with horizontal connections between the two that concatenate feature maps at different spatial scales. We added these components to the 3D

U-Net individually to devise four new neural network architectures (Residual U-Net, Inception U-Net, Dense U-Net, and Squeeze-and-Excitation U-Net). The components were added to each convolutional layer within each spatial scale and did not carry past the consequent downsampling or upsampling layers. The rectified linear unit (ReLu) activation was used in all layers, with the exception of the final sigmoid output. Batch normalization was applied after each convolutional layer for regularization. Our networks were implemented in DeepNeuro with Keras/TensorFlow backend.[10]

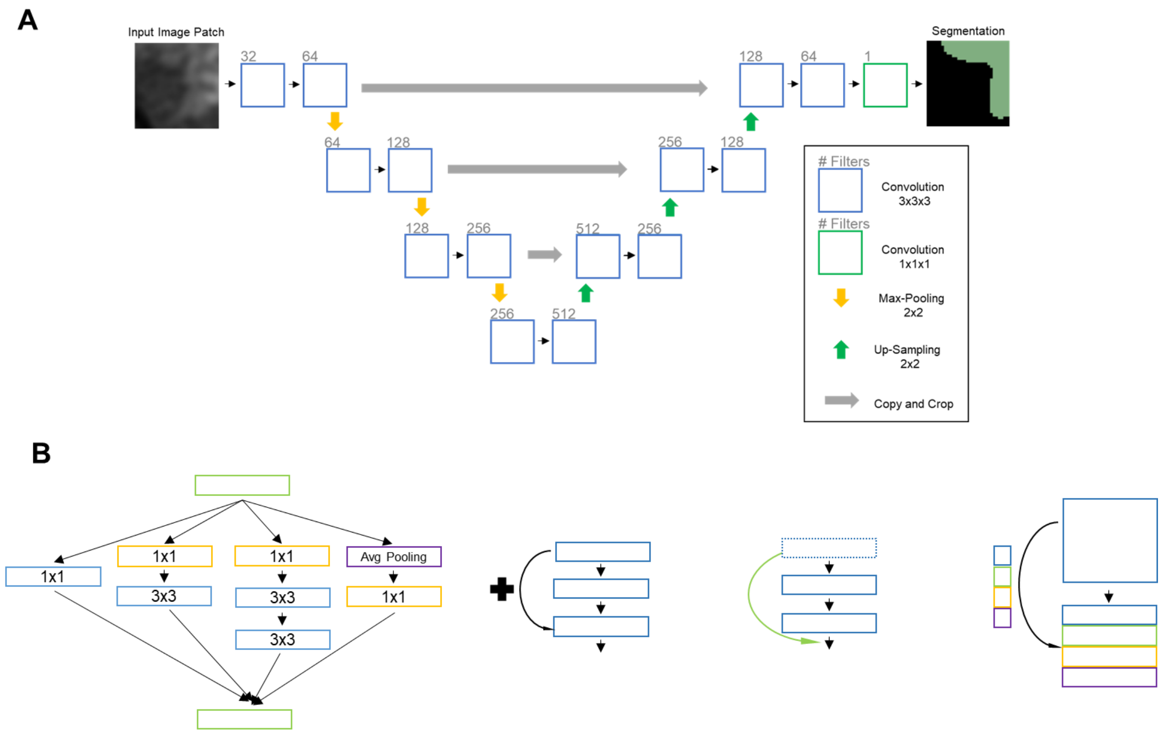


Figure 6-2. U-Net architecture (A) was modified with (B) inception, residual, dense, and squeeze-and-excitation modules.

The network was trained iteratively through all extracted patches on a NVIDIA Tesla P100 graphics processing unit. Binary segmentation maps were generated by binarizing the probability maps at a threshold of .5. We used Nestorov Adaptive Moment Estimation (Nadam) to train the 3D U-Nets with an initial learning rate 10^{-5}, minimizing a soft Dice loss function[315]:

$$(1)\ \ D(q,p) = \frac{2\sum_i p_i q_i}{\sum_i (p_i + q_i) + \alpha}$$

where D is Sørensen–Dice coefficient, q is the probability output of the neural network, p is the ground truth, and α is a smoothing constant set to 1 in our experiments. Twenty patches of size 64x64x8 mm voxels were extracted for each patient in the training set. Two channels were used, one for the b0 DWI image and one for the b1000 DWI image. The chosen patch size provided enough image context for segmentation while still being computationally and memory efficient. Patches were extracted from non-ischemic and ischemic lesions in a 1:1 ratio. To prevent overfitting and to increase the size of the training set, patches were augmented by means of sagittal flips.[137] Four patches were extracted for each patient in the Validation Set and the soft dice was evaluated at the end of each training epoch. Training was stopped when Validation Set soft dice did not improve for ten consecutive epochs. Once the network was trained, inference of new DWI images was performed by inputting successive patches of size 62x62x6, with neighboring patches having an overlap ratio of 15/16. The smaller patch size and overlap criteria at inference time was used to mitigate any edge effects. Probability maps for each of these patches were then predicted by the model, and voxels with predictions from multiple overlapping patches had their probabilities averaged.

Because averaging the output of multiple trained machine learning models has been shown to improve performance, the performance of model ensembles was also evaluated.[261,316] The improved performance from ensembling is analogous to how a consensus of experts is more likely to be correct than any single expert.[223] The Top 2, Top 3, and Top 4 models with different neural network architectures based on Testing Set Dice Similarity Coefficient as well as all 5 models were ensembled by averaging the output probability maps. The averaged probability maps were then binarized at a threshold of .5. The performance of ensembling multiple models of the same neural network architecture was also assessed – to do so, we trained 3 additional Inception U-Nets and assessed the performance of an ensemble of 2, 3, and 4 Inception U-Nets.

6.3.5 Qualitative assessment by stroke neurologist and two radiologists

To qualitatively assess the quality of the stroke segmentations, 94 segmentations annotated by either manually or the algorithm were randomly selected. We then automatically determined the axial slice with the largest lesion area. The segmentations were then overlaid on the b1000 DWI images and assessed by a stroke neurologist (Rater 1), neuroradiologist (Rater 2), and radiologist (Rater 3) blinded to whether the segmentations were performed manually or automatically. Specifically, each rater was asked to answer 3 questions: 1) Would you edit the segmentation? 2) What is the quality of the segmentation on a scale of 1-4? 3) Was it a human annotator (as opposed to the algorithm)?

6.3.6 Statistical analysis

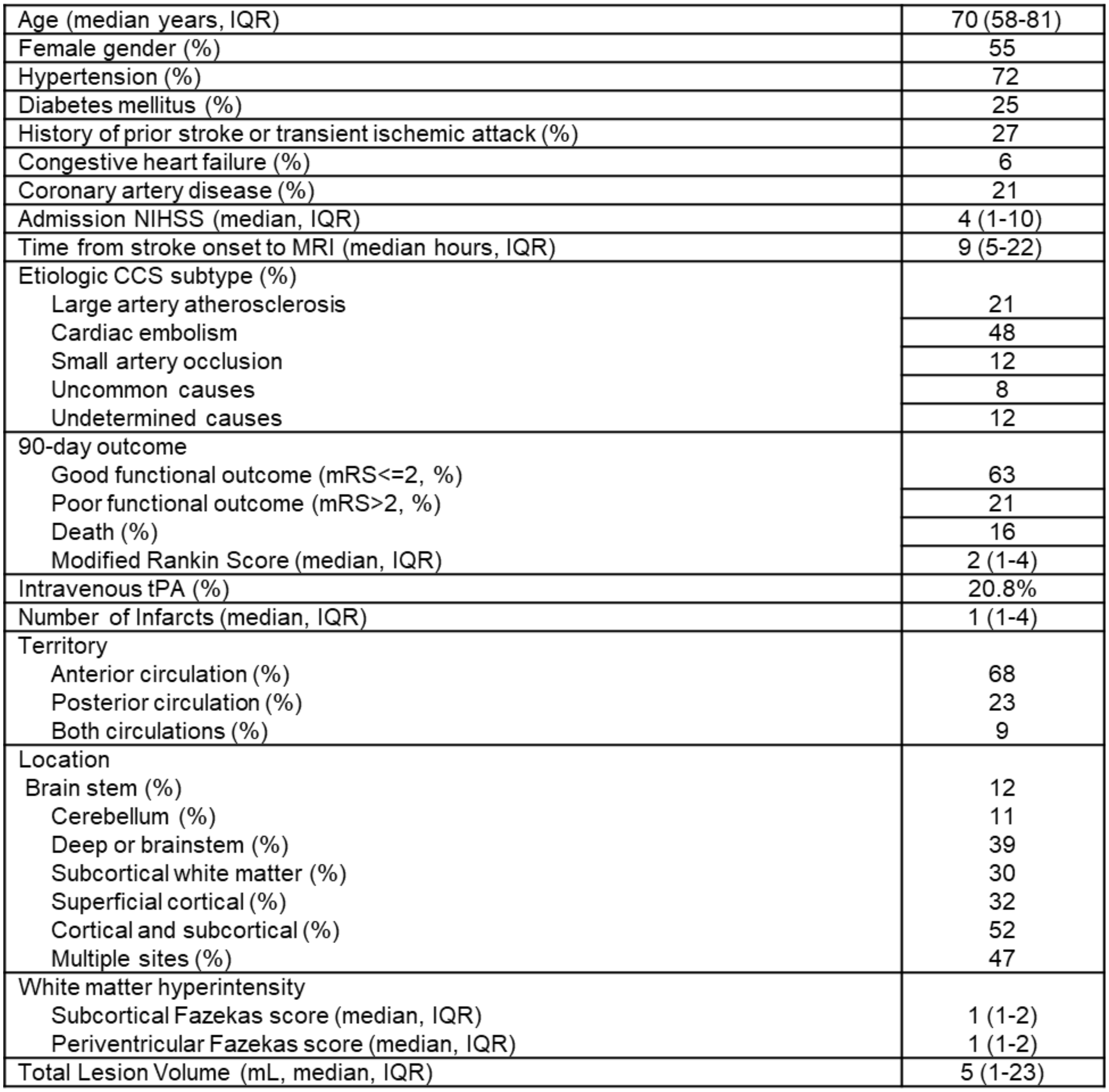
The performance of individual models and model ensembles were evaluated by means of Testing Set Sørensen–Dice Similarity Coefficient with the Mann–Whitney U test. We assessed clinical and imaging stroke features that determined model performance by examining differences between first and last quartiles of Dice Similarity Coefficient. Clinical and image stroke features that were significant on univariate analysis (Mann–Whitney U test, $p < .05$) were inputted into a multivariate logistic regression to assess for significance as threshold of $p = .05$. Additionally, we assessed Dice Similarity Coefficient for patients with small and large infarct volumes, defined as patients with manual volumes below and above the median manual volume for all patients, respectively. Spearman's rank correlation coefficient (ρ) was used to evaluate the relationship between dice coefficient, manual volume, and time to MR imaging (from stroke onset). We also calculated the stroke detection rate, defined as the percentage of patients with at least one true positive voxel (a voxel that was segmented as ischemic by both the rater and the algorithm).[121] Relatedness between volumes derived from manual and automatic segmentations was assessed via Intraclass Correlation Coefficient from one-way analysis of variance (R version 3.1.2). To compare similarity between manual and automatic segmentations, we utilize the Chi-squared test (Questions 1 and 3) and Mann-Whitney U test (Question 2). We also evaluated manual vs automatic volumes for patients stratified by the mRS at 90-days after admission for ischemic stroke.[313] We compared volumes for mRS <= 2 and > 2 (excluding mRS values of 6), which represents 90-day functional outcome. We also compared mRS <=5 and 6, which represents 90-day survival. The volumes between the different mRS stratifications were evaluated using the Mann–Whitney U test. The threshold for significance for all statistical tests was $p = 0.05$.

6.4 Results

6.4.1 Patient cohort

The study cohort comprised of 1205 consecutive patients with acute ischemic stroke. Patient demographics and stroke characteristics are shown in Table 6-1. The study population reflected typical characteristics of a consecutive hospital population with ischemic stroke; the median age was 70 and median admission NIHSS score of 4. The population also reflected the heterogeneity in stroke characteristics; 68% of patients had anterior circulation stroke, 12%

presented with a lacunar infarct, and 47% had multiple infarcts on MRI. Median time from symptom onset to DWI was 9 hours. There was a total of 5142 infarcts in 1205 patients, with infarct volumes falling into a wide range of values from .004 to 818.120 mL.



Age (median years, IQR)	70 (58-81)
Female gender (%)	55
Hypertension (%)	72
Diabetes mellitus (%)	25
History of prior stroke or transient ischemic attack (%)	27
Congestive heart failure (%)	6
Coronary artery disease (%)	21
Admission NIHSS (median, IQR)	4 (1-10)
Time from stroke onset to MRI (median hours, IQR)	9 (5-22)
Etiologic CCS subtype (%)	
Large artery atherosclerosis	21
Cardiac embolism	48
Small artery occlusion	12
Uncommon causes	8
Undetermined causes	12
90-day outcome	
Good functional outcome (mRS<=2, %)	63
Poor functional outcome (mRS>2, %)	21
Death (%)	16
Modified Rankin Score (median, IQR)	2 (1-4)
Intravenous tPA (%)	20.8%
Number of Infarcts (median, IQR)	1 (1-4)
Territory	
Anterior circulation (%)	68
Posterior circulation (%)	23
Both circulations (%)	9
Location	
Brain stem (%)	12
Cerebellum (%)	11
Deep or brainstem (%)	39
Subcortical white matter (%)	30
Superficial cortical (%)	32
Cortical and subcortical (%)	52
Multiple sites (%)	47
White matter hyperintensity	
Subcortical Fazekas score (median, IQR)	1 (1-2)
Periventricular Fazekas score (median, IQR)	1 (1-2)
Total Lesion Volume (mL, median, IQR)	5 (1-23)

Table 6-1. Patient demographic information and clinical and imaging stroke features

6.4.2 Performance of Individual Deep Learning Models for Segmentation

The performance of the five architectures (U-Net, Residual U-Net, Inception U-Net, Dense U-Net, and Squeeze-And-Excitation U-Net) was investigated on the Testing Set. The average time for segmentation was 16 seconds using our trained algorithms. The best performing individual model was the Inception U-Net, which had a median dice similarity coefficient of 0.72 (0.697-0.751) within the Testing Set (Table 6-2). Notably, the performance of Inception U-Net was better than the standard U-Net ($p < .05$) within the Testing Set.

	Training Set (n = 720)	Validation Set (n = 243)	Testing Set (n = 242)
U-Net	0.715 (0.689-0.731)	0.691 (0.658-0.731)	0.680 (0.626-0.714)
Residual U-Net	0.709 (0.685-0.734)	0.673 (0.633-0.723)	0.678 (0.615-0.707)
Inception U-Net	0.740 (0.721-0.765)	0.725 (0.693-0.751)	0.72 (0.697-0.751)
Dense U-Net	0.693 (0.666-0.717)	0.701 (0.658-0.733)	0.696 (0.643-0.718)
Squeeze-And-Excitation U-Net	0.696 (0.667-0.72)	0.667 (0.630-0.694)	0.650 (0.599-0.688)

Table 6-2. Median dice similarity coefficient (95% Confidence Interval) of individual models within the Training, Validation, and Testing Sets.

6.4.3 Performance of ensembles of different U-Net architectures

We also assessed the performance of ensembling the individual models of different U-Net architectures. The median dice similarity coefficient on the Testing Set for an Ensemble of Top 2 Models, Top 3 Models, Ensemble of Top 4 Models, and Ensemble of All 5 Models was 0.726 (0.68-0.747), 0.724 (0.682-0.753), 0.722 (0.694-0.746), and 0.71 (0.686-0.738), respectively (Table 6-3). The best performing ensemble was the Ensemble of Top 2 Models (Inception and Dense U-nets). This performance was significantly better than that of a single U-Net ($p < .05$) but not from a single Inception U-Net.

	Training Set (n = 720)	Validation Set (n = 243)	Testing Set (n = 242)
Ensemble of Top 2	0.737 (0.715-0.757)	0.733 (0.698-0.763)	0.726 (0.680-0.747)
Ensemble of Top 3	0.738 (0.717-0.753)	0.731 (0.707-0.760)	0.724 (0.682-0.753)
Ensemble of Top 4	0.737 (0.719-0.756)	0.723 (0.690-0.752)	0.722 (0.694-0.746)
Ensemble of All 5	0.733 (0.712-0.751)	0.719 (0.689-0.743)	0.71 (0.686-0.738)

Table 6-3. Median dice similarity coefficient (95% Confidence Interval) of model ensembles of different U-Net architectures within the Training, Validation, and Testing Sets.

6.4.4 Performance of ensembles of inception U-Nets

Additionally, we assessed the performance of ensembling Inception U-Nets. Within the Testing Set, the median dice similarity coefficient of the Ensemble of 2, 3, and 4 Inception U-Nets was 0.729 (0.696-0.753), 0.734 (0.708-0.75), and 0.737 (0.708-0.765) (Table 6-4). Notably, the performance of all ensembles of Inception U-Nets were higher than that of a Single Inception U-Net. The best performing ensemble was that of 4 Inception U-Nets. Example manual and automatic segmentations from the Ensemble of 4 Inception U-Nets are shown in Fig 6-3. This performance was significantly better than that of a single U-Net (p < .005) but not from a single Inception U-Net (p = .18).

	Training Set (n = 720)	Validation Set (n = 243)	Testing Set (n = 242)
Inception U-Net	0.740 (0.721-0.765)	0.725 (0.693-0.751)	0.720 (0.697-0.751)
Ensemble of 2	0.764 (0.740-0.778)	0.739 (0.702-0.774)	0.729 (0.696-0.753)
Ensemble of 3	0.758 (0.742-0.777)	0.741 (0.713-0.77)	0.734 (0.708-0.75)
Ensemble of 4	0.765 (0.749-0.781)	0.746 (0.721-0.774)	0.737 (0.708-0.765)

Table 6-4. Median dice similarity coefficient (95% Confidence Interval) of model ensembles of Inception U-Nets within the Training, Validation, and Testing Sets.

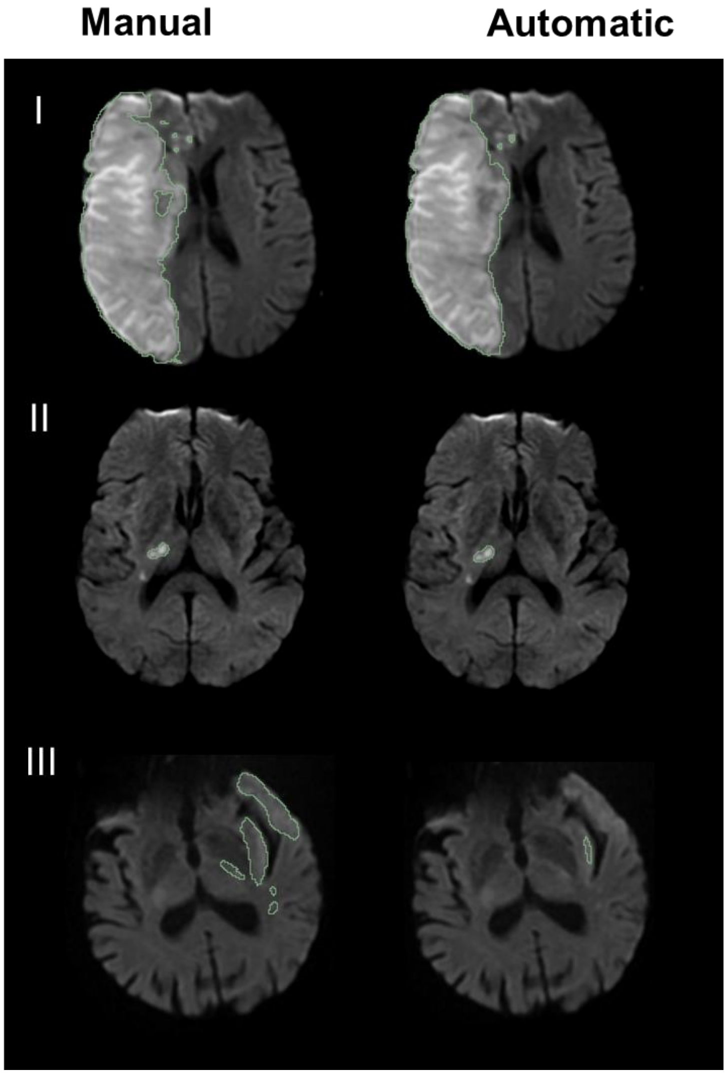


Figure 6-3. Example of manual vs automatic segmentations showing high (I and II) and low (III) agreement.

Within the Validation and Testing Sets, the Ensemble of 4 Inception U-Nets had a stroke detection rate of 92.8%. The average volume of lesions detected was 25.011 mL while the average volume of lesions missed was 0.082 mL (Fig. 6-4). For patients with small infarcts (<median), the

median dice coefficient was 0.657 (0.588-0.701). For patients with large infarcts, the median dice coefficient was 0.816 (0.795-0.829). There was a moderate association between dice coefficient and manual volume (Spearman’s ρ = 0.561, p < .001, Fig. 6-5). There was no association between dice coefficient and time from symptom onset to MR imaging (ρ = 0.097) as well as between manual volume and time to MR imaging (ρ = 0.158). The intraclass correlation coefficient between manually and automatically derived infarct volumes (from ensemble of 4 Inception U-Nets) was 0.977 (p <. 0001) within in the Validation and Testing Sets (Fig. 6-6). Example automatic segmentations are shown in Fig. 4. In case III, the automated algorithm missed the parts of the infarction in inferior frontal and temporal regions that are typically subject to susceptibility artifacts.

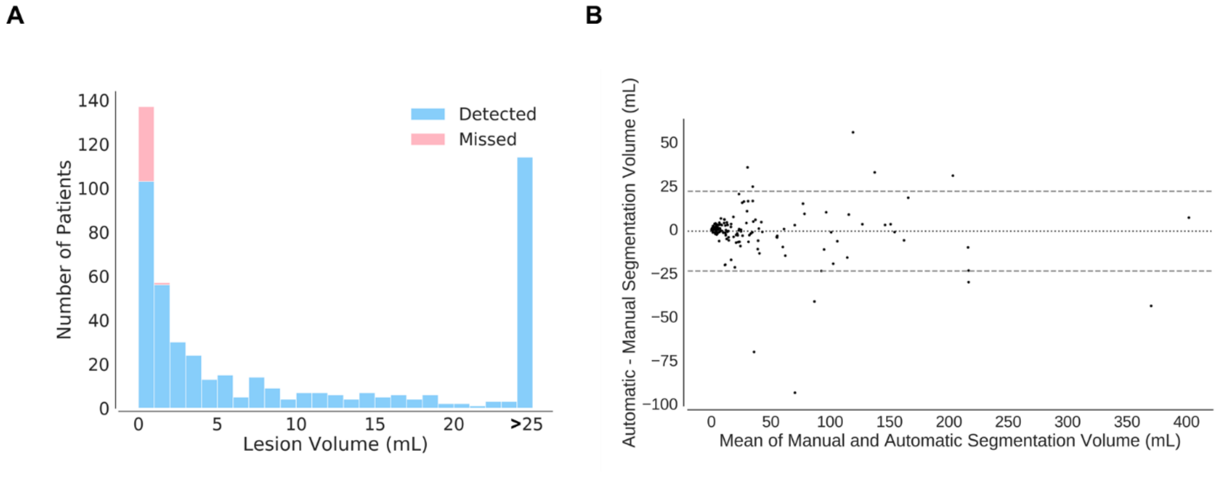


Figure 6-4. Histogram of manual lesion volumes for (A) detected and missed lesions. (B) Bland-Altman plot.

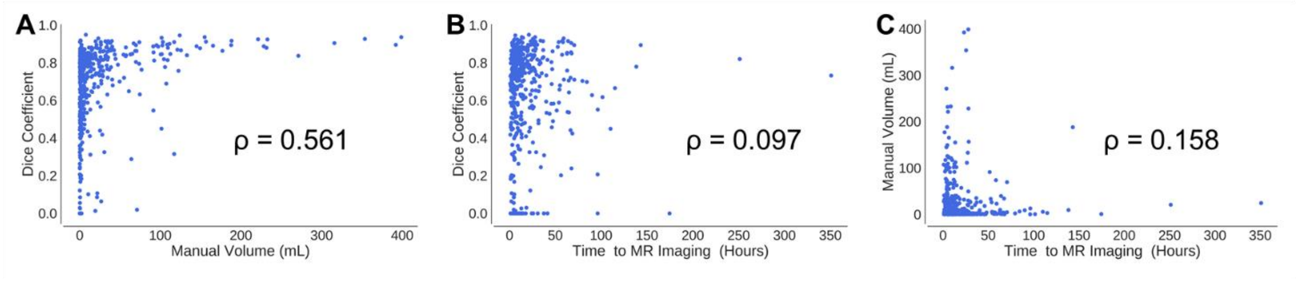


Figure 6-5. Scatter plot of (A) dice coefficient vs manual volume, (B) dice coefficient vs time to MR imaging, and (C) manual volume vs time to MR imaging.

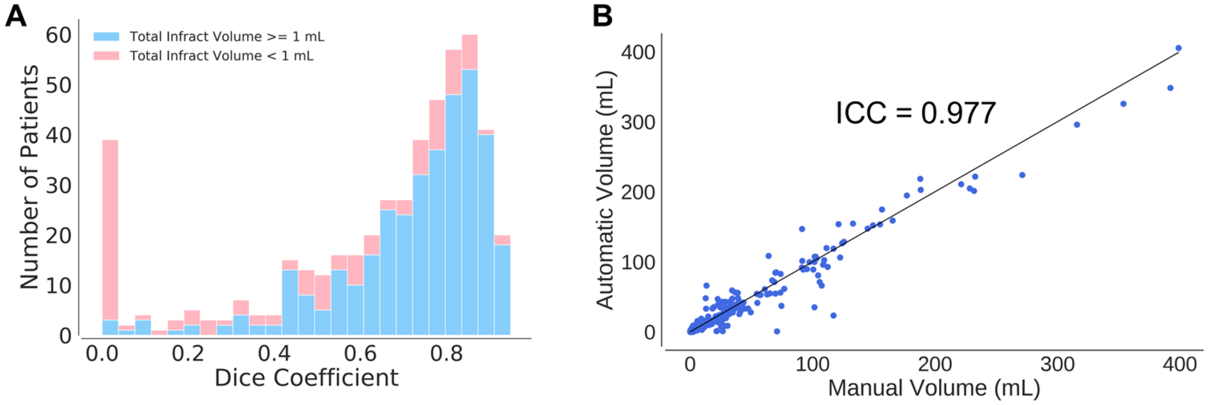


Figure 6-6. (A) Histogram of Dice. (B) Scatter plot of automatic vs manual volumes.

Despite high intraclass correlation coefficient, there were 17 patients (7%) in the Testing dataset where dice similarity index was equal to zero, meaning that the automated algorithm completely missed the manually outlined lesion. A retrospective analysis of such cases revealed that all 17 had a punctate infarct measuring < 1 mL (Fig. 6-7). Bivariate analyses comparing the first and last quartiles of dice similarity coefficients revealed that infarct size, admission stroke severity as measured by NIHSS score, anterior territory, infarct location (isolated deep or brainstem, isolated superficial cortical, cortical and subcortical, and multiple sites) predicted the algorithms performance (Table 6-5). A multivariate analysis showed that only infarct size (p <

0 .01), isolated deep or brainstem location (p < .05), and isolated cortical location (p < 0.05) were the independent predictors of performance.

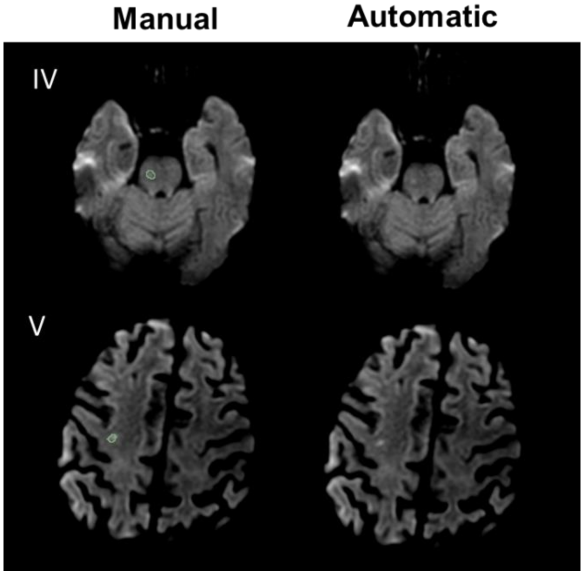
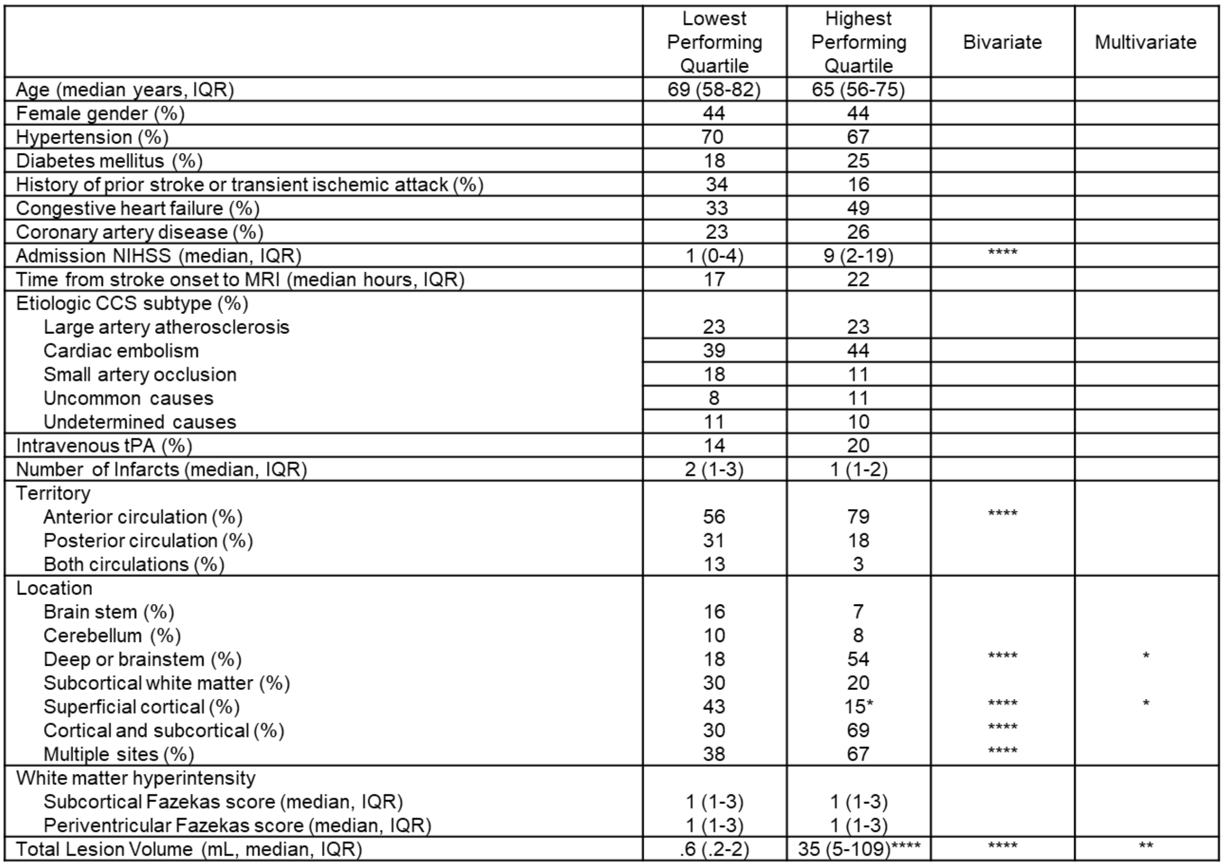


Figure 6-7. Examples of two cases < 1 mL that were missed by DeLTIS.



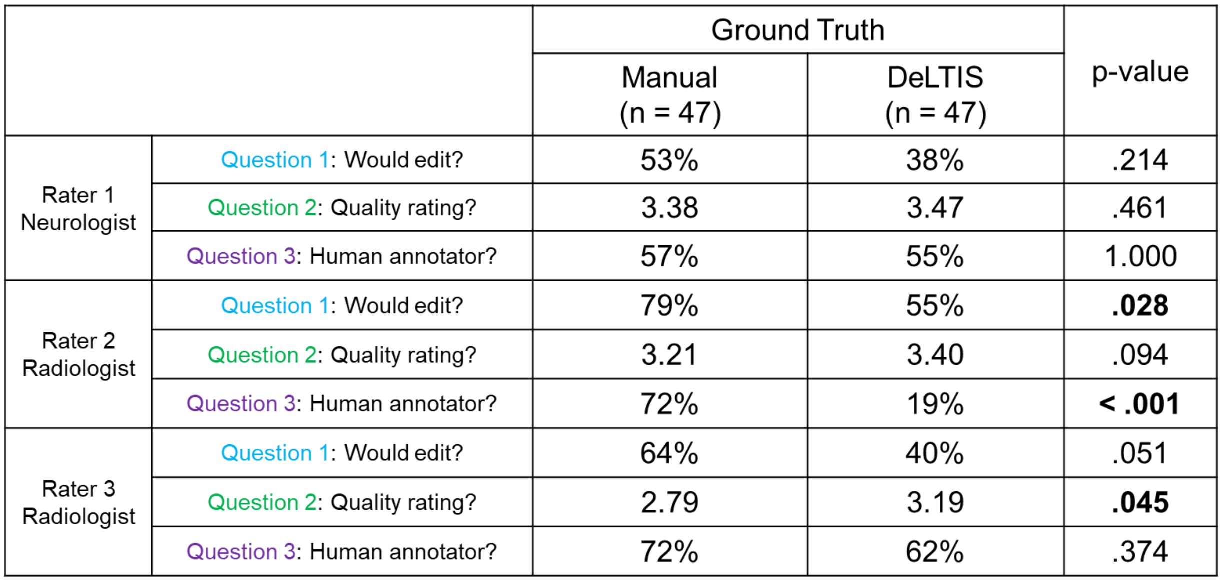
	Lowest Performing Quartile	Highest Performing Quartile	Bivariate	Multivariate
Age (median years, IQR)	69 (58-82)	65 (56-75)		
Female gender (%)	44	44		
Hypertension (%)	70	67		
Diabetes mellitus (%)	18	25		
History of prior stroke or transient ischemic attack (%)	34	16		
Congestive heart failure (%)	33	49		
Coronary artery disease (%)	23	26		
Admission NIHSS (median, IQR)	1 (0-4)	9 (2-19)	****	
Time from stroke onset to MRI (median hours, IQR)	17	22		
Etiologic CCS subtype (%)				
Large artery atherosclerosis	23	23		
Cardiac embolism	39	44		
Small artery occlusion	18	11		
Uncommon causes	8	11		
Undetermined causes	11	10		
Intravenous tPA (%)	14	20		
Number of Infarcts (median, IQR)	2 (1-3)	1 (1-2)		
Territory				
Anterior circulation (%)	56	79	****	
Posterior circulation (%)	31	18		
Both circulations (%)	13	3		
Location				
Brain stem (%)	16	7		
Cerebellum (%)	10	8		
Deep or brainstem (%)	18	54	****	*
Subcortical white matter (%)	30	20		
Superficial cortical (%)	43	15*	****	*
Cortical and subcortical (%)	30	69	****	
Multiple sites (%)	38	67	****	
White matter hyperintensity				
Subcortical Fazekas score (median, IQR)	1 (1-3)	1 (1-3)		
Periventricular Fazekas score (median, IQR)	1 (1-3)	1 (1-3)		
Total Lesion Volume (mL, median, IQR)	.6 (.2-2)	35 (5-109)****	****	**

Table 6-5. Bivariate and multivariate analyses comparing the first and last quartiles of dice similarity coefficients. * p < .05, ** p <.01, **** p < .001

6.4.5 Qualitative assessment by clinical raters

There were no statistically significant differences between manual and DeLTIS segmentations for Questions 1-3 for Rater 1 (Table 6-6). For Rater 2, there were no statistically significant differences between manual and DeLTIS segmentations for Question 2. However, Rater 2 would have edited 79% of the manual segmentations as opposed to 55% of the DeLTIS segmentations (p < .05). Additionally, Rater 2 believed 72% of the manual segmentations were performed by humans and 19% of the DeLTIS segmentations were performed by humans (p < .001). For Rater 3, there were no statistically significant differences between manual and

DeLTIS segmentations for Questions 1 and 3. However, Rater 3 rated DeLTIS segmentations of higher quality than manual segmentations (p < .05).



		Ground Truth		p-value
		Manual (n = 47)	DeLTIS (n = 47)	
Rater 1 Neurologist	Question 1: Would edit?	53%	38%	.214
	Question 2: Quality rating?	3.38	3.47	.461
	Question 3: Human annotator?	57%	55%	1.000
Rater 2 Radiologist	Question 1: Would edit?	79%	55%	**.028**
	Question 2: Quality rating?	3.21	3.40	.094
	Question 3: Human annotator?	72%	19%	**< .001**
Rater 3 Radiologist	Question 1: Would edit?	64%	40%	.051
	Question 2: Quality rating?	2.79	3.19	**.045**
	Question 3: Human annotator?	72%	62%	.374

Table 6-6. Qualitative assessment of 47 manual and 47 DeLTIS segmentations by an stroke neurologist (Rater 1), neuroradiologist (Rater 2), and radiologist (Rater 3). Question 1 asked whether the rater would edit the segmentation. Question 2 asked the rater to grade the quality of the segmentation on a scale of 1-4. Question 3 asked whether the rater believed the segmentation was performed by a human (as opposed to an algorithm). Results from question 1 and 3 are shown as percent of cases the rater stated yes. Results from question 2 are shown as the mean rating.

6.4.6 Manual and automatic volume by 90-day mRS score

90-day mRS was missing for 71 patients and these patients were excluded from the analysis. Patients were stratified based on mRS score at <= 2 vs > 2 (representing 90-day functional outcome) and at <= 5 vs 6 (representing 90-day survival). Within the Validation and

Testing Sets, the median manually derived volumes were 2.21 (1.87-2.75) mL, 9.79 (5.93-18.20) mL, 2.97 (2.43-3.69) mm^3, and 38.79 (27.97-76.69) mL for patients with a 90-day mRS score of <= 2, >2, <=5, and 6, respectively. The median of automatically derived volumes from the Ensemble of 4 Inception U-Nets was 1.96 (1.62-2.52) mL, 13.60 (5.25-18.82) mL, 2.86 (2.16-3.66) mL, and 41.44 (25.30-56.30) mL, respectively. For the manually derived volumes, there was a statistically significant difference between patients with mRS score <= 2 vs. > 2 (p<.001) and mRS score <=5 vs. >5 (p<.001). Similarly, for the automatically derived volumes, there was a statistically significant difference between patients with mRS score <= 2 and > 2 (p<.001) and mRS score <=5 vs. >5 (p<.001) (Fig. 6-8).

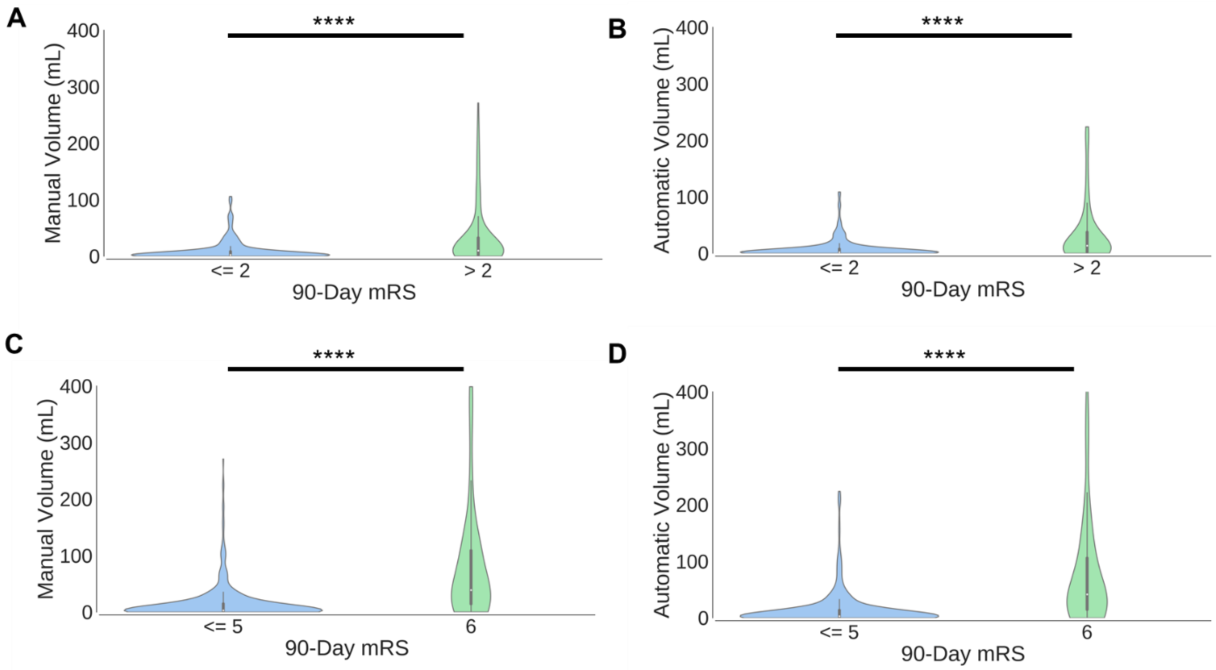


Figure 6-8. Violin plots of manual and automatic volumes for disability (A-B) and survival (C-D). ****p<.001

6.5 Discussion

In this section, we demonstrate the utility of DeLTIS, a fully automated, deep-learning based pipeline for segmentation of acute infarcts on DWI, as part of a larger effort to apply deep learning techniques to the field of neurology. We show that with minimal image pre-processing (i.e. no resampling and brain extraction) along with a 3D U-Net architecture, our automated pipeline achieves high performance in the Testing Set with a median dice of 0.737. In addition, DeLTIS offers a high stroke detection rate, at 92.8% on the Validation and Testing Sets. We further show that, qualitatively, automatic segmentations are the same or superior to manual segmentations via ratings from three raters; all raters stated that they would edit a greater proportion of the manual segmentations compared to the automatic segmentations and one of the raters graded automatic segmentations as of significantly higher quality than manual segmentation. In addition to segmentation accuracy, we also evaluated infarct volumes as derived from manual segmentations and automatic segmentations, and found that they showed high agreement. Finally, we demonstrate that automatically segmented volumes confer comparable predictive information to manual volumes for 90-day functional outcome and mortality. Hence, the automatically generated segmentations can be used instead of manual segmentations in stroke research examining quantitative infarct metrics and clinical end points.

Infarct volume measurements are becoming an integral piece of stroke research. Continuous nature of the infarct volume data allows for exploring associations in smaller samples and making inferences with fewer data points as compared to categorical assessments based on visual inspection of neuroimaging. Also, categorical classifications suffer from high interrater disagreement. For instance, the interrater agreement to determine whether infarct size is less than or greater than one-third of the middle cerebral artery territory, which roughly corresponds to 100 ml, is only moderate (kappa = 0.4).[317] Infarct volume information is also frequently used by clinicians in practice for prediction of tissue and clinical outcome,[318,319] assessment of the risk of developing hemorrhagic transformation or malignant edema,[320,321] and assessment of eligibility for thrombolytic treatment or endovascular thrombectomy. Most clinical trials of intravenous and endovascular recanalization therapies have excluded patients who have already developed large infarcts because the risk of treatment complications such as symptomatic intracerebral hemorrhage outweighs the anticipated benefit in large infarcts.[322] Infarcts exceeding one third of the middle cerebral artery territory are considered to be contraindication for intravenous thrombolysis.[322] Similarly, most endovascular thrombectomy protocols exclude patients based on certain infarct volume thresholds that range from 20-70 ml depending on other associated clinical and imaging features of stroke. Conversely, some protocols attempt to avoid exposing patients with small infarcts to the risks, discomfort, and cost associated with recanalization treatments as such small infarcts causing minute structural brain injury in the absence of large vessel occlusion confer a high recovery potential regardless of treatment. The major premise of the present study is that it provides a rapid and accurate means of obtaining infarct volume data; our automated algorithm provides infarct volumes within seconds. In contrast, manual outlining can take anywhere from a few minutes to half an hour depending on the lesion load and the experience level of the operator. Furthermore, in patients with multiple scattered infarcts, manual outlining takes even more time. In our experience, average time required to manually outline a patient’s infarcts hovers around 15 minutes for expert neuroradiologists. DeLTIS can generate lesion outlines rapidly and with minimal level of inconsistency and thus could be particularly useful in settings where there are large quantities of data, such as in big consortia and multicenter repositories.

The use of a large dataset allowed us to examine the performance characteristics of DeLTIS in different patient populations. We found that infarct size and infarct location were the independent predictors of the algorithm's performance; DeLTIS performed less well in patients with punctate infarcts (median volume = 0.155 mL) especially when such lesions were located in deep white matter or within the cortex. Detection of such small dots by clinicians are often prompted by the accompanying signs and symptoms. While clinical information is unarguable important in lesion detection, it should be noted that even with clinical data, agreement between human experts for detection of such punctate dots is typically very low, leading to a potential attribution bias.[310] Automated tools like DeLTIS are agnostic to clinical information and thus provide segmentations with limited accuracy but without attribution bias. It is possible to further enhance the accuracy of automated tools by incorporating clinical information in the future.

We studied a large and unselected patient population with diverse mechanisms of stroke, brain morphology, and radiographic stroke features such as infarct size, shape, and signal intensity. Hence, DeLTIS effectively accounts for the variance in patient populations and infarct characteristics, which are key for its generalizability. We showed that the performance of automated segmentations improved upon the 3D U-Net architecture with the addition of Inception modules as well as ensembling multiple Inception U-Nets. The improved performance of the Inception U-Net is likely due to multi-scale learning allowed by the multiple pathways with different convolution filter sizes.[144] Similarly, the improved performance of ensembling is due to the decreased probability of overfitting when using the consensus output of multiple trained models. Because our approach is designed to distinguish abnormality from normal brain anatomy, it has potential utility for other structural neurological abnormalities, such as hemorrhage, tumors, and plaques.

6.6 Limitations

Despite such strengths, there are several possible improvements to this section as well. First, manual volumetric segmentations were derived from a single rater for each patient. Future studies can evaluate the performance of DeLTIS compared to manual segmentations by multiple raters. Secondly, we applied a basic image augmentation technique (sagittal flipping) to increase the size of training set and improve segmentation performance. The incorporation of additional image augmentation techniques, such as the use of generative adversarial networks, may bring further gains in performance.[323] Additionally, future algorithms that incorporate structural and perfusion MR could improve segmentations on DWI . Also, measures of uncertainty could be added to our pipeline to flag segmentations that require further verification from clinicians.[262,324,325] This would allow for more reliable integration into clinical workflows. While the addition of structural and perfusion MRI might increase segmentation performance, the use of multimodal imaging can comprise the method when there is a defect in acquisition. DeLTIS is solely based on DWI because DWI is sensitive to early infarcts, provides excellent conspicuity for lesion identification, and is routinely obtained in most stroke practices. Moreover, DWI alone reduces the amount of imaging that needs to be performed on a patient. Nonetheless, there are limitations of relying on a single modality as well; DWI lesions can become larger or smaller depending on how the lesion is windowed. In the present study, the reader was allowed to window the images, as they would be in real practice, to be able to detect subtle changes. In contrast, DeLTIS used a default window setting generated by the scanner. This might have caused DeLTIS to miss some punctate lesions with subtle signal intensity. Finally, although time from stroke onset to imaging was not an independent predictor of performance, DWI signal intensity can increase as a function of time during the hyperacute phase of stroke, and this might have blurred the observed correlations between manual and automatic segmentations.

6.7 Background on viscerotoxic brain infarcts

Classic textbook teaching has been that problems in internal organs, such as atrial fibrillation in the heart, lead to problems in the brain, such as ischemic stroke. However, emerging evidence suggests that acute brain injury could independently lead to internal organ injury as well, often with serious outcomes ranging from transient dysfunction to permanent morphological injury in internal organ systems. This form of injury, which is called neurogenic organ injury (NOI), is thought to result from excessive activation of or withdrawal of inhibitory inputs on central autonomic modulation centers by stroke lesions resulting in pathologically increased activity of the autonomic nervous system.[326,327] While autonomic response is generally considered systemic, i.e., response throughout the system is total, organic brain injury can cause organ-selective activation where manifestations depend on the organ involved.[328,329] Organ specificity may indicate the existence of a viscerotopic organization in the brain, analogous to the somatotopic organization, where each organ or visceral function is governed by discrete regions of the brain. It is currently not fully known what parts of the human brain predispose to NOI when injured. Accurate identification of such brain regions is critical to recognize patients at risk of NOI. Nonetheless, animal models are too limited to test the neuroanatomic hypotheses in a piecemeal way. Classic lesion-based paradigms for neuroanatomic localization in humans have been useful to suggest hypotheses but since they are based on a priori anatomical assumptions, they lack precision especially when the region of interest is large. Unprejudiced localization requires a mapping technique that takes brain as a whole rather than focusing on a particular region of interest. In this section, we sought to identify the neuroanatomic correlates of a broad range of cardiac and systemic alterations occurring after ischemic stroke using a method that is free from the bias of an a priori hypothesis as to any specific location. Our goal was to understand how internal organ dysfunction after acute ischemic stroke might be mediated. For this, we tested the hypothesis that there are brain regions which, when infarcted, are associated with NOI in a large and prospective cohort without primary causes of internal organ injury. We further aimed to understand the direction of the relationship between infarct location and NOI and quantify the added burden of infarct location on clinical outcome.

6.8 Methods

6.8.1 Patient cohort

We explored the neuroanatomic correlates of four different post-stroke cardiac or systemic abnormalities (CSA) that included plasma cardiac troponin T (cTnT) elevation as a marker of structural cardiac injury, QT segment prolongation on ECG as a marker of electrophysiological cardiac alteration, pneumonia and urinary tract infection (UTI) as a marker of altered pulmonary, urinary, or immune system functioning, and acute stress hyperglycemia (ASH) as a marker of increased glycogenolysis in the liver. A corrected QT interval (QTc) was calculated using the Bazett's formula: QTc = QT interval / square root of the RR interval (sec). For 1208 patients in the prospective, longitudinal, consecutive, NIH-funded study (Heart-Brain Interactions Study), a neuroradiologist manually generated binary maps of acute infarcts on DWI.

6.8.2 Image processing and statistical analysis

The diffusion weighted images and corresponding outline images were co-registered to the Montreal Neurological Institute (MNI) 152 template using a 12-degrees of freedom affine transformation via the BRAINSFit module in 3D Slicer.[258,330] This was followed by iterative groupwise elastic registration using SimpleElastix.[331] The diffusion images and corresponding outline images were subsequently re-sampled at 4 mm isotropic resolution in for faster permutation calculation. P-value maps were generated using threshold-free cluster enhancement via Randomise in FMRIB Software Library with sex and age as covariates (Fig. 6-9A).[332,333] Using a nonparametric permutation test with 5,000 permutations, significance was reported at a family-wise error corrected $p < .05$.[334] We also generated two additional neuroanatomic maps in cohorts with cTnT elevation and QTc prolongation who were initially excluded due to the presence of a non-neurogenic cause for the CSA. These maps aimed to serve logic check purpose to further support the existence of a neurogenic link. Overlap of each patient with the resulting neuroanatomic maps was calculated. Logistic regression models were fit with an overlap ratio of 10% for each of the neuroanatomic maps to determine the odds ratio for an abnormal lab test, 90-day disability (90-day modified Rankin Score > 2), and 90-day survival, correcting for infarct volume.

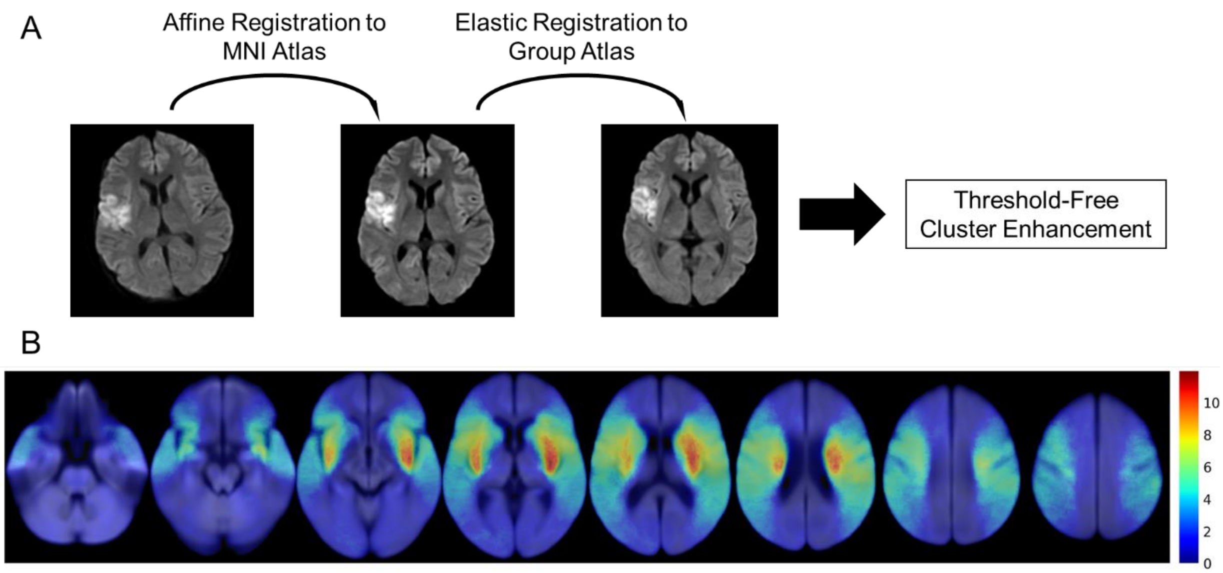

A
Affine Registration to MNI Atlas
Elastic Registration to Group Atlas
Threshold-Free Cluster Enhancement
B
10
8
6
4
2
0

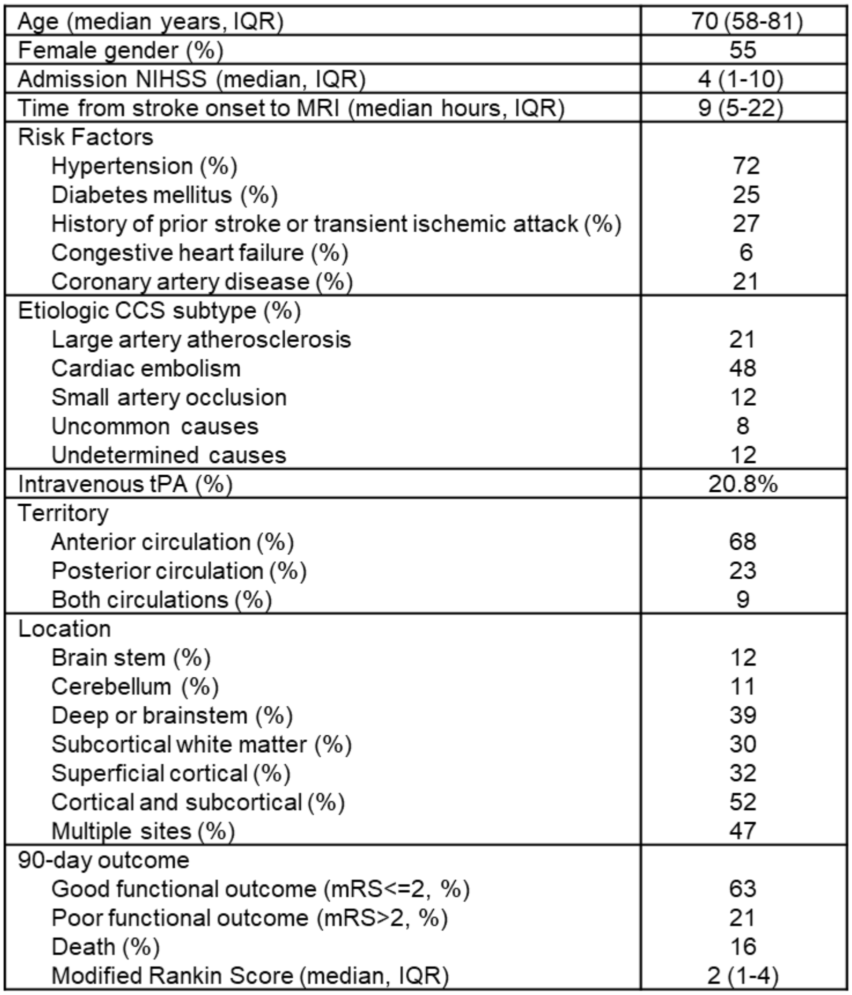
Figure 6-9 (A) To bring each patient image into the same space, affine registration to the MNI Atlas was applied followed by groupwise elastic registration. Then threshold-free cluster enhancement via permutation was applied to reveal each viscerotopic map. (B) Heatmap of ischemic infarcts across all patients, color-coded by the % of patients that have a lesion at a given voxel.

6.9 Results

6.9.1 Patient cohort

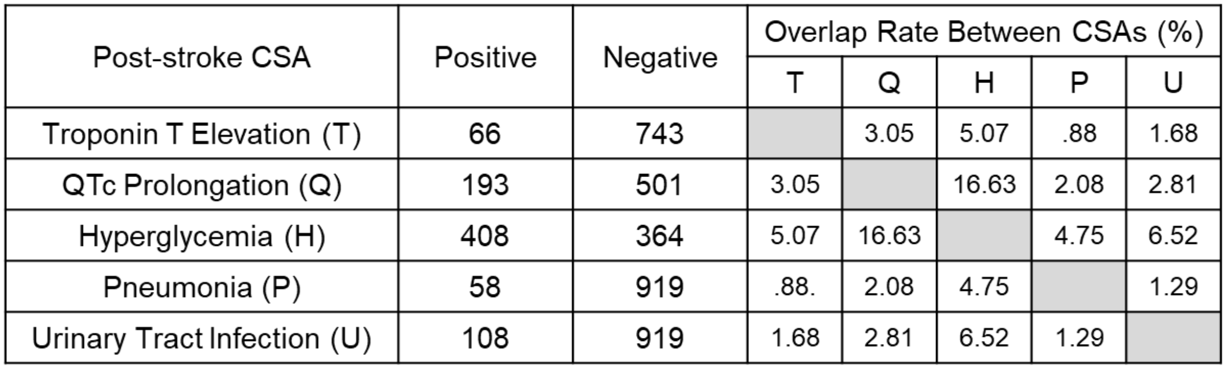
We screened a total of 1474 consecutive patients with ischemic stroke admitted within 72 hours of stroke onset during the study period. We excluded 159 patients in whom an MRI could not be obtained because of contraindications. We excluded an additional 59 patients who underwent an MRI study in an outside hospital and images were not available for analysis. Of the remaining 1256 patients, we excluded 48 patients because of extensive motion artifacts in MRI. Hence, the final study population consisted of 1208 patients. Table 6-7 shows baseline characteristics and clinical and imaging stroke features of the study population. Fig. 6-9B demonstrates topographic distribution of coregistered binary infarct maps showing infarct probability in all 1208 consecutive patients. The burden of stroke on the brain was mainly on deep hemispheric gray and white matter structures.

Table 6-8 illustrates the recruitment process into each study. The cTnT substudy comprised of 813 patients, of whom 66 (8.1%) had elevated cTnT levels that could not be attributed to a known cause. The QTc substudy included 694 patients, of whom 194 (27.8%) had QTc prolongation in the absence of a known provoker. The ASH substudy included 772 patients, of whom 408 (52.8%) met the criteria for diagnosis of ASH. The pneumonia substudy included 977 patients, 58 (5.9%) of which has pneumonia. The UTI substudy included 1027 patients, 108 of which had UTI (10.5%). The overlap rates between the different CSAs are shown in Tables 6-8 and 6-9.



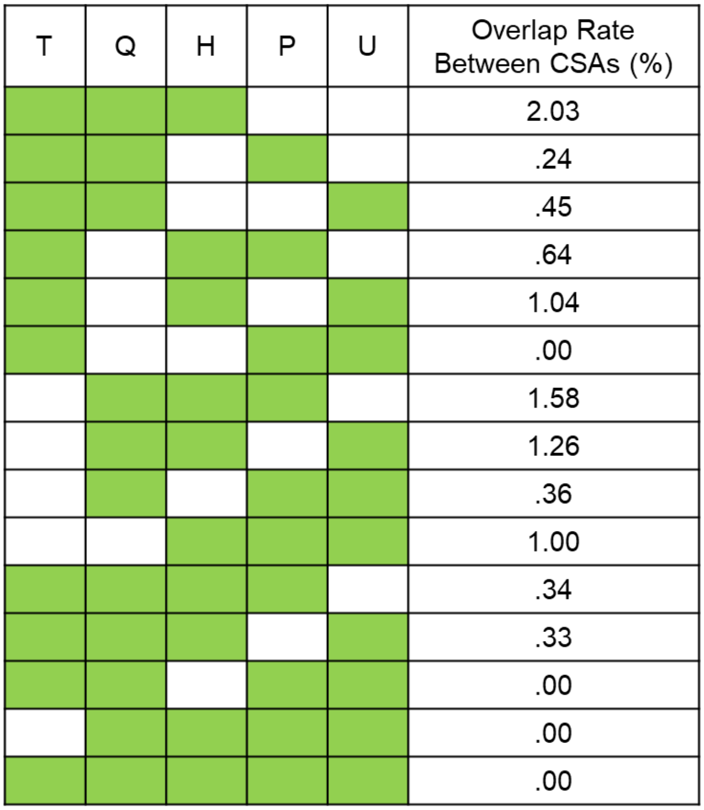
Age (median years, IQR)	70 (58-81)
Female gender (%)	55
Admission NIHSS (median, IQR)	4 (1-10)
Time from stroke onset to MRI (median hours, IQR)	9 (5-22)
Risk Factors	
Hypertension (%)	72
Diabetes mellitus (%)	25
History of prior stroke or transient ischemic attack (%)	27
Congestive heart failure (%)	6
Coronary artery disease (%)	21
Etiologic CCS subtype (%)	
Large artery atherosclerosis	21
Cardiac embolism	48
Small artery occlusion	12
Uncommon causes	8
Undetermined causes	12
Intravenous tPA (%)	20.8%
Territory	
Anterior circulation (%)	68
Posterior circulation (%)	23
Both circulations (%)	9
Location	
Brain stem (%)	12
Cerebellum (%)	11
Deep or brainstem (%)	39
Subcortical white matter (%)	30
Superficial cortical (%)	32
Cortical and subcortical (%)	52
Multiple sites (%)	47
90-day outcome	
Good functional outcome (mRS<=2, %)	63
Poor functional outcome (mRS>2, %)	21
Death (%)	16
Modified Rankin Score (median, IQR)	2 (1-4)

Table 6-7. Patient demographic information and clinical and imaging stroke features



Post-stroke CSA	Positive	Negative	Overlap Rate Between CSAs (%)				
			T	Q	H	P	U
Troponin T Elevation (T)	66	743		3.05	5.07	.88	1.68
QTc Prolongation (Q)	193	501	3.05		16.63	2.08	2.81
Hyperglycemia (H)	408	364	5.07	16.63		4.75	6.52
Pneumonia (P)	58	919	.88.	2.08	4.75		1.29
Urinary Tract Infection (U)	108	919	1.68	2.81	6.52	1.29	

Table 6-8. Number of positive and negative cases with abnormal lab values as well as percentage binary positive overlap with other abnormal lab values (relative to the total number of patients who received the two lab value tests).

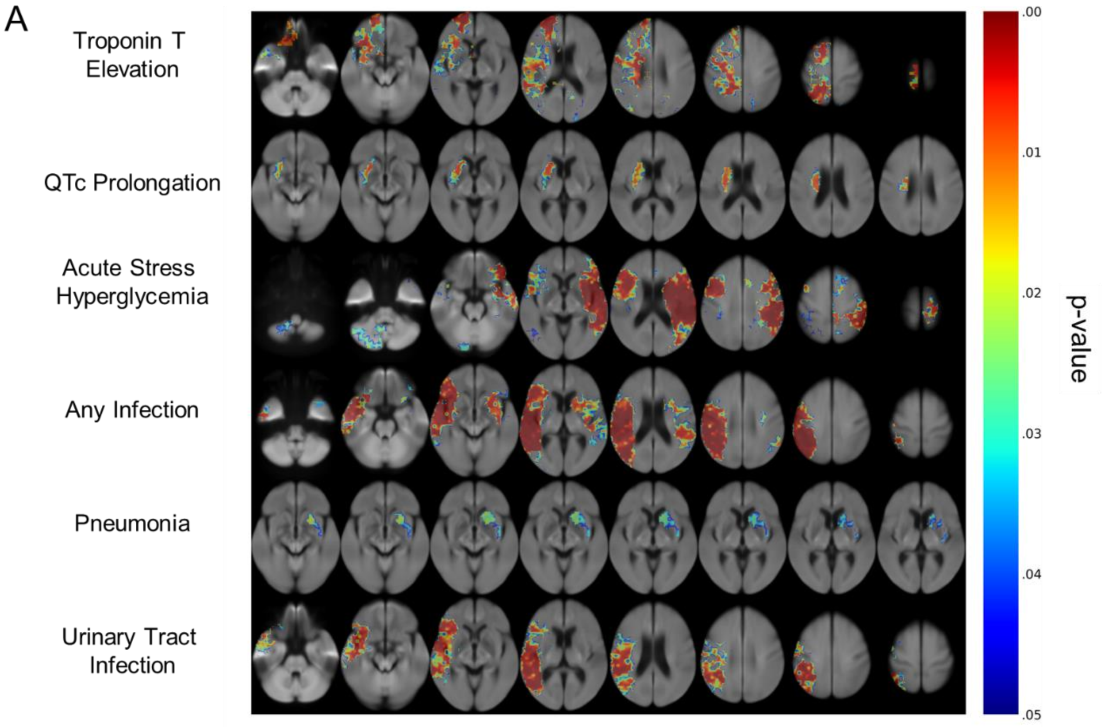


T	Q	H	P	U	Overlap Rate Between CSAs (%)
🟩	🟩	🟩			2.03
🟩	🟩		🟩		.24
🟩	🟩			🟩	.45
🟩		🟩	🟩		.64
🟩		🟩		🟩	1.04
🟩			🟩	🟩	.00
	🟩	🟩	🟩		1.58
	🟩	🟩		🟩	1.26
	🟩		🟩	🟩	.36
		🟩	🟩	🟩	1.00
🟩	🟩	🟩	🟩		.34
🟩	🟩	🟩		🟩	.33
🟩	🟩		🟩	🟩	.00
	🟩	🟩	🟩	🟩	.00
🟩	🟩	🟩	🟩	🟩	.00

Table 6-9. Percentage positive tertiary/quaternary/quinary overlap with other abnormal lab values (relative to the total number of patients who received the three/four/five lab value tests) for Troponin T Elevation (T), QTc Prolongation (Q), Hyperglycemia (H), Pneumonia (P), and Urinary Tract Infection (U).

6.9.2 Neuroanatomic maps for each cardiac or systemic abnormality

Fig. 6-10 demonstrates the neuroanatomic maps for each CSA. We identified at least one cluster for each CSA. The clusters for cTnT elevation and QTc prolongation were both located in the right hemisphere. There were three clusters for ASH, a small cluster in the right hemisphere, a small cluster in the cerebellum, and a large one in the left hemisphere. There were two clusters for post-stroke infection, one in the right and the second one in the left hemisphere. Infection type specific maps revealed that the cluster on the left was exclusively associated with pneumonia whereas the one on the right with UTI. All CSA maps displayed overlap with the insula and opercula regions (Fig. 6-11, Table 6-10). Pneumonia had the largest percentage of the neuroanatomical map that was located in the insula. ASH had the largest percentage of the neuroanatomical map that was located in the opercula. In contrast, no statistically significant maps neuroanatomical maps were revealed for cTnT elevation and QTc prolongation logic checks (Fig. 6-12).



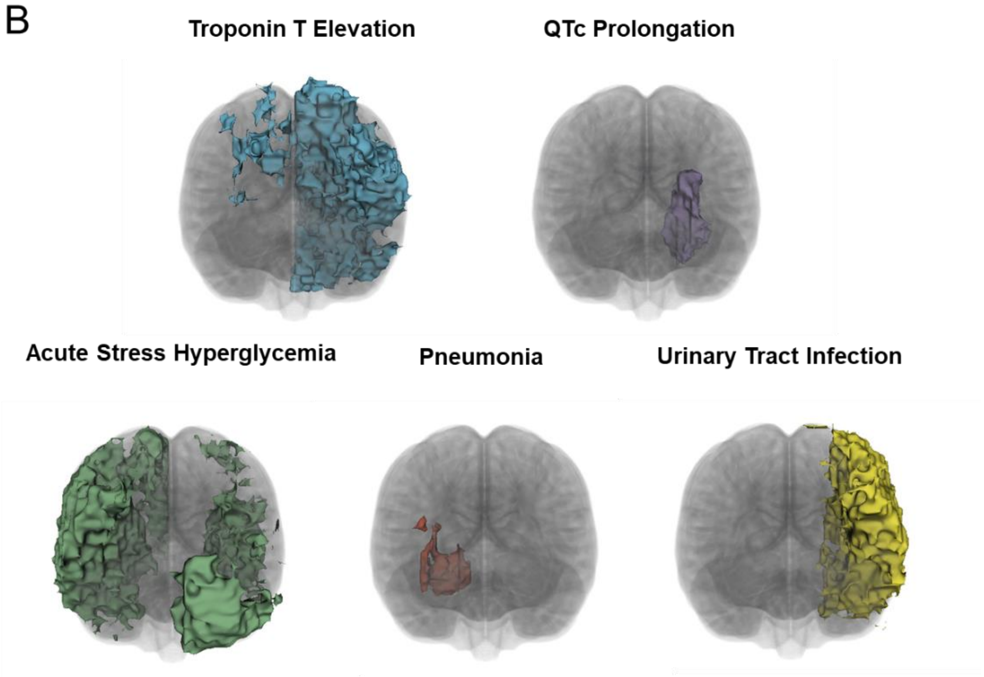


Figure 6-10. (A) Viscerotopic maps revealed from cluster analysis, color-coded by p-value and overlaid on the groupwise DWI atlas (L-R radiological convention) (B) 3D models of viscerotopic maps (L-R non-radiological convention).

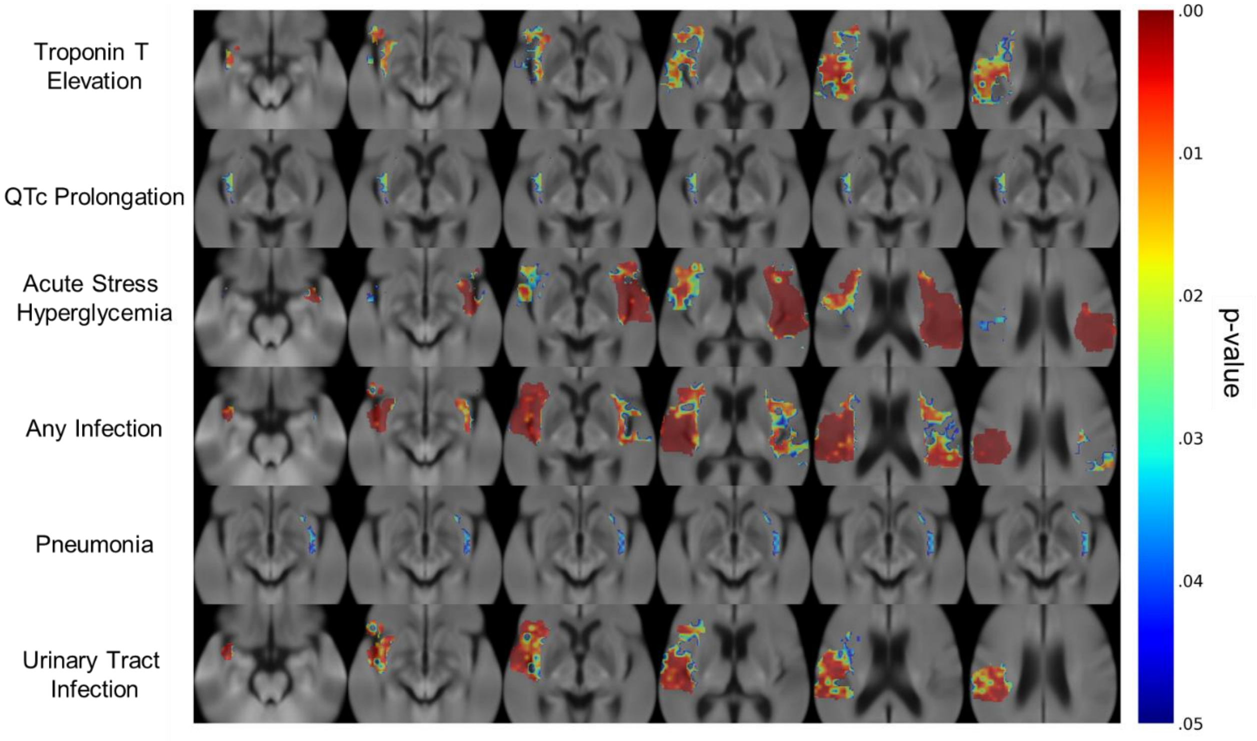


Figure 6-11. The component of the viscerotopic maps within the insula and opercula, color-coded by p-value and overlaid on the groupwise DWI atlas.

Post-stroke CSA	**Percentage of the Map that Encompassed**		
	Insula %	**Opercula %**	**Other %**
Troponin T Elevation	2.8	13.2	84.0
QTc Prolongation	8.4	3.8	87.8
Acute Stress Hyperglycemia	3.7	24.2	72.1
Any Infection	5.5	26.9	67.6
Pneumonia	19.6	8.8	71.6
Urinary Tract Infection	3.7	28.4	67.9

Table 6-10. Anatomical description of maps in terms of percentage of the map that is in the insula, opercula, and other brain regions.

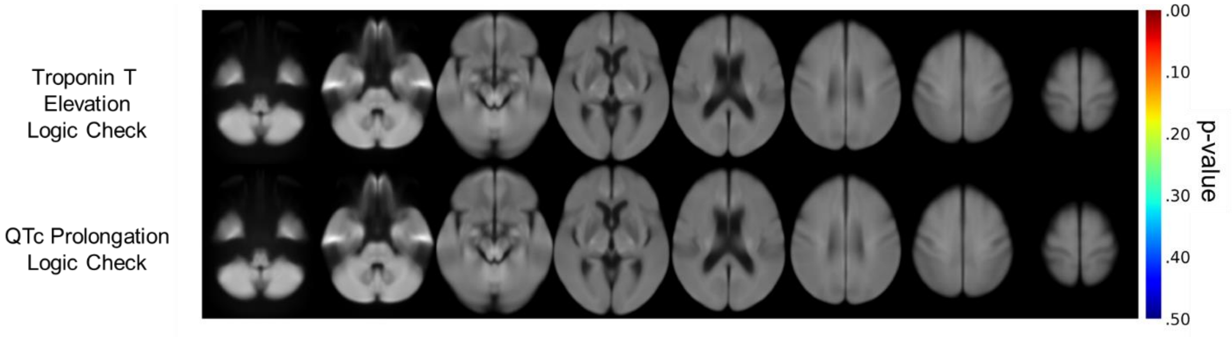

Troponin T Elevation Logic Check
QTc Prolongation Logic Check
.00
.10
.20
.30
.40
.50
p-value

Figure 6-12. Logic check maps for Troponin T Elevation and QTc Prolongation, color-coded by p-value and overlaid on the groupwise DWI atlas. There were no voxels with $p < .05$ for either logic check map.

6.9.3 Predictive value of neuroanatomical maps

Overlap with all maps were predictive of post-stroke CSA (Table 6-11, Figure 6-13). Overlap with QTc prolongation and pneumonia maps was predictive of 90-day functional disability. Overlap with the pneumonia map was predictive of 90-day mortality. When looking only at insula/opercula regions within the neuroanatomical maps, overlap with those regions was predictive of QTc prolongation and pneumonia. Overlap with the neuroanatomic pneumonia map within the insula/opercula regions was also predictive of 90-day mortality (Table 6-12).

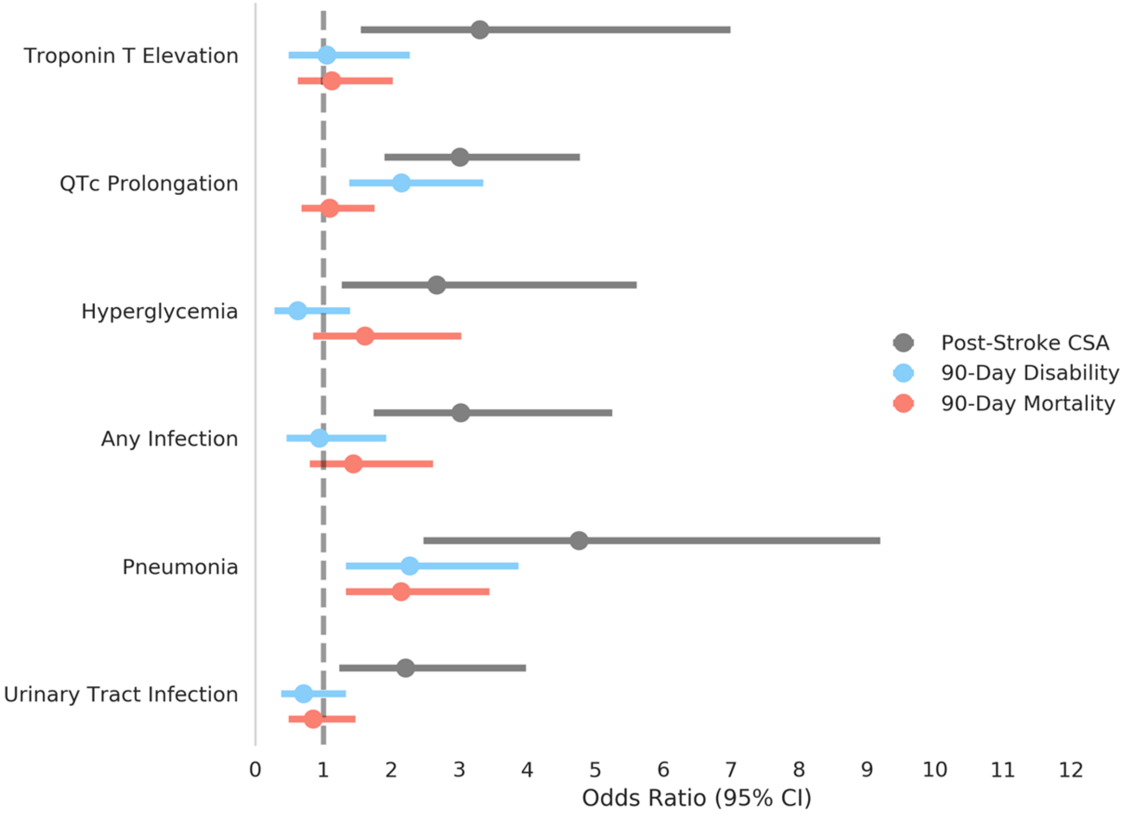
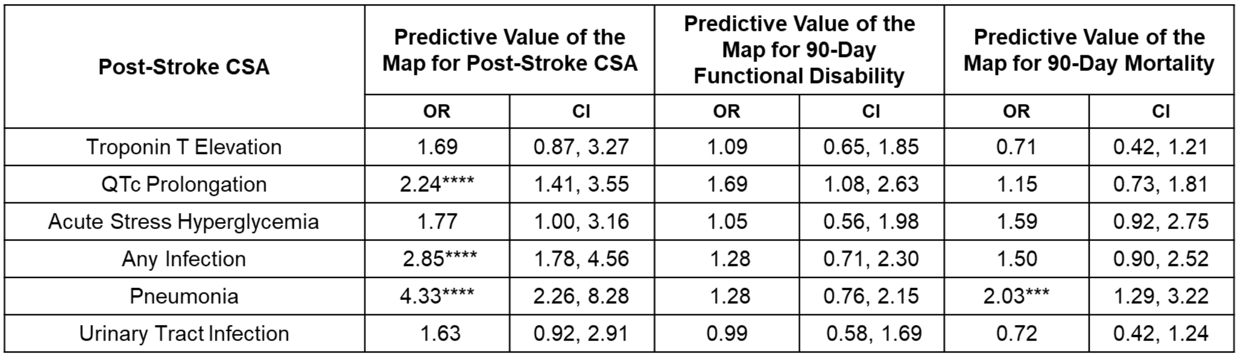


Figure 6-13. The odds ratio (OR) and 95% confidence interval (CI) of developing abnormal test results, 90-day functional disability, and 90-day mortality with a 10% overlap threshold after adjusting for infarct volume.

Post-Stroke CSA	Predictive Value of the Map for Post-Stroke CSA		Predictive Value of the Map for 90-Day Functional Disability		Predictive Value of the Map for 90-Day Mortality	
	OR	CI	OR	CI	OR	CI
Troponin T Elevation	3.30***	1.55, 6.99	1.05	0.49, 2.27	1.12	0.62, 2.02
QTc Prolongation	3.01****	1.90, 4.77	2.15****	1.38, 3.35	1.09	0.68, 1.75
Acute Stress Hyperglycemia	2.67**	1.27, 5.61	0.62	0.28, 1.39	1.61	0.85, 3.03
Any Infection	3.02****	1.74, 5.25	0.94	0.46, 1.92	1.44	0.80, 2.61
Pneumonia	4.76****	2.47, 9.19	2.27***	1.33, 3.87	2.14***	1.33, 3.44
Urinary Tract Infection	2.21**	1.23, 3.98	0.71	0.38, 1.33	0.85	0.49, 1.47

Table 6-11. The odds ratio (OR) and 95% confidence interval (CI) of developing abnormal test results, 90-day functional disability, and 90-day mortality with a 10% overlap threshold after adjusting for infarct volume. *p < .05, **p < .01, ***p < .005, ****p < .001



Post-Stroke CSA	Predictive Value of the Map for Post-Stroke CSA		Predictive Value of the Map for 90-Day Functional Disability		Predictive Value of the Map for 90-Day Mortality	
	OR	CI	OR	CI	OR	CI
Troponin T Elevation	1.69	0.87, 3.27	1.09	0.65, 1.85	0.71	0.42, 1.21
QTc Prolongation	2.24****	1.41, 3.55	1.69	1.08, 2.63	1.15	0.73, 1.81
Acute Stress Hyperglycemia	1.77	1.00, 3.16	1.05	0.56, 1.98	1.59	0.92, 2.75
Any Infection	2.85****	1.78, 4.56	1.28	0.71, 2.30	1.50	0.90, 2.52
Pneumonia	4.33****	2.26, 8.28	1.28	0.76, 2.15	2.03***	1.29, 3.22
Urinary Tract Infection	1.63	0.92, 2.91	0.99	0.58, 1.69	0.72	0.42, 1.24

Table 6-12. The odds ratio (OR) and 95% confidence interval (CI) of developing abnormal test results, 90-day functional disability, and 90-day mortality with a 10% overlap threshold in the insulas and opercula sub-region after adjusting for infarct volume. *p < .05, **p < .01, ***p < .005, ****p < .001

6.10 Discussion

In this section, we provide support for the principle of NOI, showing that the location of acute brain injury is correlated with symptoms that the patient can later develop, specifically cTnT elevation, QT segment prolongation, acute stress hyperglycemia, pneumonia, and UTI. Notably, we show that adjustment for infarct volume did not alter the relationship between infarct location and the CSAs suggesting that infarct location confers independent predictive information for developing post-stroke CSAs. Additionally, we show that the location dependence for cTnT elevation, QT segment prolongation was abolished when patients with non-neurogenic causes of symptoms were included, serving as a logic check for these neuroanatomical maps. The fact that neuroanatomical maps for cTnT elevation and QT segment prolongation were on the right side of the brain also provide further support for NOI as well, given that the brain is, on the most part, organized contralaterally with respect to the body.

It is important to note that all maps had overlap with the insula and opercula, regions of the brain that are known to have homeostatic functionality. Interestingly, the neuroanatomic map for pneumonia localized on the left side of the brain while the map for UTI localized on the right side of the brain, which may correspond a currently unknown mechanism of NOI.

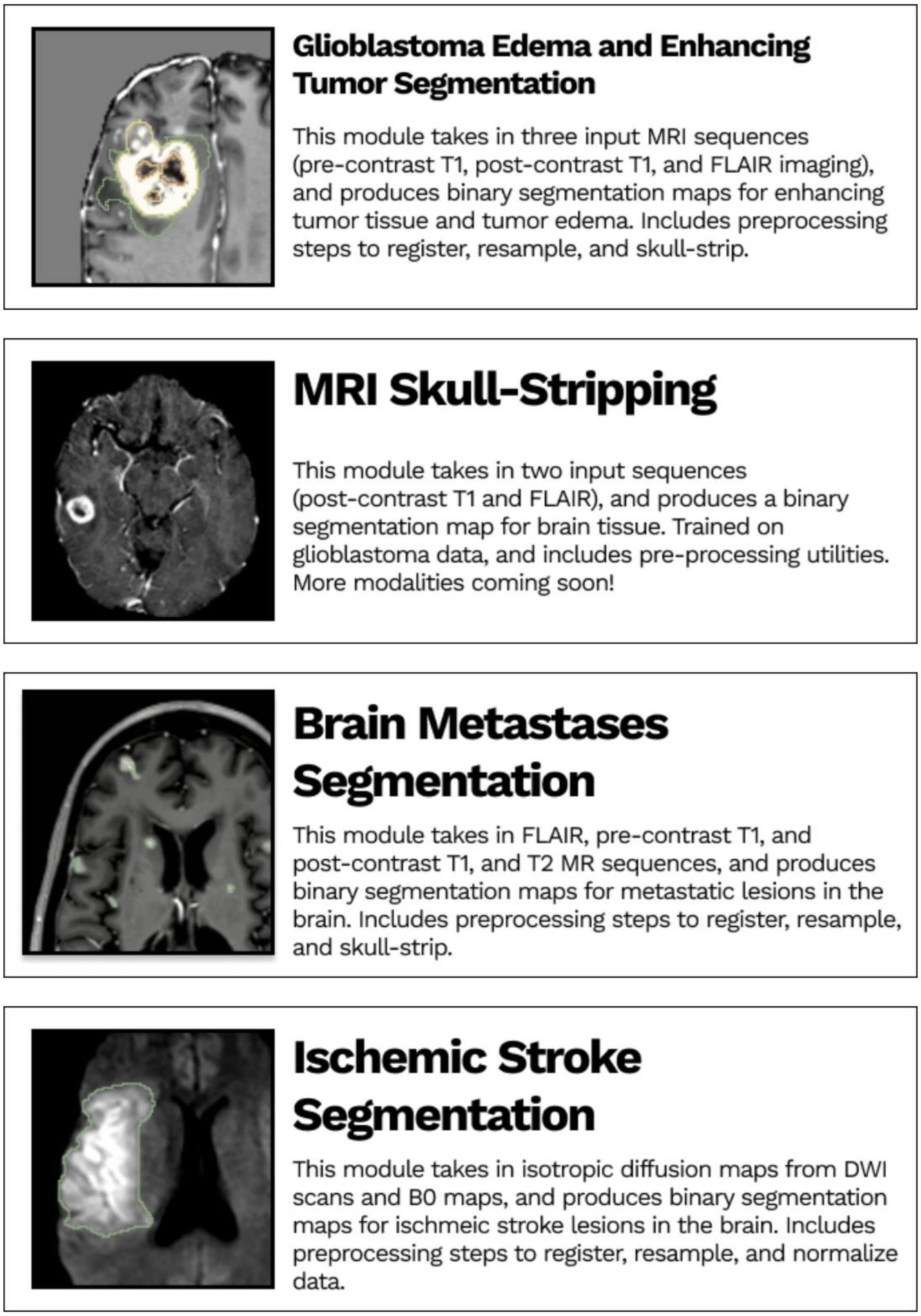
We also show that overlap with the neuroanatomic maps for QTc prolongation as well as pneumonia provide prognostic information with the patient. This can serve as a potentially useful tool for clinicians at the time of imaging for clinical decision-making, although further prospective study is needed.

6.11 Limitations

There are three main limitations to this work. First, the neuroanatomical maps were generated using manual segmentations from experts, which may be subject to inter-rater variability. Future studies will use automatic segmentation (such as DelTIS described earlier in this chapter) for enhanced reproducibility of the neuroanatomical maps. Second, once the neuroanatomical maps were generated, overlap with the maps was determined as a fraction of overlap, which was subsequently used to calculate odds ratios of developing the symptom or for patient prognosis. An extension of this would be to treat the maps not as binary masks but rather probabilistically as certain neuroanatomic regions within each map were more significant than others. Alternatively, the probabilistic neuroanatomic map can be combined with a given patient’s lesion outline into another neural network that is tasked with predicting the symptom risk or prognosis. Lastly, in our interpretation of the neuroanatomical maps, we focused only on the insula and opercula, brain regions known to regular homeostasis. Further analysis can assess other brain regions for their role in NOI.

6.12 Open-source deep learning for neuroimaging

One major challenge within deep learning is reproducibility.[335] Oftentimes, the methods section of papers is incomplete and subtle differences in pre-processing, deep learning hyperparameters, and post-processing can compromise the performance of resulting models. Furthermore, different versions of software packages may also change performance. As such, we developed an open-source deep learning package for neuroimaging, DeepNeuro (https://github.com/QTIM-Lab/DeepNeuro).[10] DeepNeuro is designed in such a way that a user with minimal coding experiencing can develop end-to-end deep learning pipelines. Furthermore, DeepNeuro is customizable with the modular design so users can modify pipeline components as needed for the same applications as our labs or their own healthcare applications. Furthermore, we have made the glioma segmentation and ischemic stroke segmentation (DeLTIS) pipelines including the final trained model publicly available in a dockerized solution. To use our pipelines, the user simply has to download the docker container from Docker Hub and run the container from command line.



Glioblastoma Edema and Enhancing Tumor Segmentation

This module takes in three input MRI sequences (pre-contrast T1, post-contrast T1, and FLAIR imaging), and produces binary segmentation maps for enhancing tumor tissue and tumor edema. Includes preprocessing steps to register, resample, and skull-strip.

MRI Skull-Stripping

This module takes in two input sequences (post-contrast T1 and FLAIR), and produces a binary segmentation map for brain tissue. Trained on glioblastoma data, and includes pre-processing utilities. More modalities coming soon!

Brain Metastases Segmentation

This module takes in FLAIR, pre-contrast T1, and post-contrast T1, and T2 MR sequences, and produces binary segmentation maps for metastatic lesions in the brain. Includes preprocessing steps to register, resample, and skull-strip.

Ischemic Stroke Segmentation

This module takes in isotropic diffusion maps from DWI scans and B0 maps, and produces binary segmentation maps for ischmeic stroke lesions in the brain. Includes preprocessing steps to register, resample, and normalize data.

Figure 6-14. Pipelines, including the trained models, for glioma and ischemic stroke segmentation are publicly available in DeepNeuro.

6.13 Conclusions

In this section, we demonstrate novel deep learning architectures for automatic segmentation of ischemic stroke. Furthermore, we show that automatically derived volumes showed high agreement with one another. Additionally, we show that automatic volumetrics can be used to stratify functional outcomes for stroke patients. Our fully-automatic pipeline for stroke segmentation demonstrates the potential for deep learning-based tools to find clinical and research utility in stroke. DELTIS presents an attractive alternative to commercially available softwares, which can be prohibitively expensive for widespread clinical use, particularly in places with limited resources.

We also used a mapping technique that is free from the bias of a-priori hypothesis as to any specific location, we show that both cardiac and systemic abnormalities occurring after stroke map to specific regions in the brain. We show that maps for all abnormalities overlap in part with the insula and opercula. We also show that these maps are predictive of the abnormalities as well as patient outcomes, showing the potential utility of the maps to aid with clinical-decision making.

7 Summary and Conclusion

We end with a summary of our methods to enhance medical imaging workflows with deep learning and a broader discussion of implications and future work. Although there are many emerging studies that have show the potential of deep learning for clinical workflows, these algorithms have yet to improve routine clinical practice. We are only beginning to see clinical trials for AI algorithms within the literature, due to the numerous challenges that prevent their clinical translation. These challenges are the central focus of this dissertation. In chapter 3, we showcase how deep learning models can be trained without sharing patient data. We found that distributed deep learning methods can be as effective as centrally hosted data. In chapter 4, we explored how various data, training, and model parameters that can influence model performance. We found that the design of algorithms can have profound downstream implications, considerations that an engineer may not have in mind when disconnected from the patients seen in clinical practice. Our results provide evidence for synergy between technical and clinical teams as algorithm design and clinical impact are inevitably intertwined. In chapter 5, we showcase an integrated pipeline for glioma, a challenging disease with a dismal prognosis. We show how deep learning can be used for detection, segmentation, and molecular marker prediction in a single pipeline, beyond single function algorithms often seen within the literature. In chapter 6, we showcase another computational pipeline, this time for ischemic stroke. Again, showing how a multi-faceted pipeline can be a multitool in the pocket of the clinician.

As anyone in research knows, scientific inquiry and advancement is never complete. When you answer one research question, five new ones pop up. This dissertation is no exception. Within distributed learning, it is still unclear how different distributed learning methods compare with one another head-to-head. Addressing heterogeneity across institutions remains a challenging problem with no clear solution. Also, providing the most stringent protections to patient privacy introduces tradeoff to model performance that need to be rigorously evaluated. In looking at implications of model design, the parameters that we looked at only represent the tip of the iceberg of possible parameters. As new deep learning algorithms continue to be developed, there are many new parameters that will need to be thoroughly interrogated. Issues related to catastrophic forgetting and model generalizability warrant further investigation as well. Within our glioma pipeline, there are still many modes that can be added to expand functionality, such as prediction of molecular markers beyond *IDH* and drug response. As new advanced imaging modalities become utilized more in regular clinical practice, they will also be integrated into this pipeline. Similarly, the ischemic stroke pipeline can be augmented with further functionality and advanced imaging modalities.

Lastly, algorithms need to be rigorously tested “in the wild”. The controlled environments in which most research takes place is not representative at worst and overly optimistic at best. To this end, we have made progress towards packaging our algorithms into usable software for deployment. However, this is only the first step. The next is much more herculean effort – achieving FDA approval and deploying the software into the hospital system on a large scale. The role of software and infrastructure engineers cannot be emphasized enough. The creation of a truly robust algorithm must also be one that is “living”, that is that is must be constantly adapting to its environment and improving. The feedback between a clinically deployed algorithm and the scientists who can make improvements will be critical.

Stepping back, this work has given me an appreciation of the necessity of an ecosystem of scientists, engineers, and physicians. To move forward, true synergy will be needed, which can only be achieved through collaboration and trust. The hope for AI is immense – to the address the shortcomings of human interpretation and decision-making. Not only that, but also to free up physician time for tasks that machines are not quite ready to tackle – the humanistic and social aspects of patient care. As this ecosystem co-evolves with integration and feedback, we can make progress toward this goal.

References

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst*. 2012:1-9. doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007
2. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process Mag*. 2012;29(6):82-97. doi:10.1109/MSP.2012.2205597
3. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proc 25th Int Conf Mach Learn*. 2008:160-167. doi:10.1145/1390156.1390177
4. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539
5. Chang K, Balachandar N, Lam C, et al. Distributed deep learning networks among institutions for medical imaging. *J Am Med Informatics Assoc*. March 2018. doi:10.1093/jamia/ocy017
6. Chang K, Balachandar N, Lam CK, et al. Institutionally Distributed Deep Learning Networks. September 2017. http://arxiv.org/abs/1709.05929. Accessed November 15, 2017.
7. Balachandar N, Chang K, Kalpathy-Cramer J, Rubin DL. Accounting for Data Variability in Multi-Institutional Distributed Deep Learning for Medical Imaging. *J Am Med Informatics Assoc*. 2020. doi:10.1093/jamia/ocaa017
8. Beers A, Chang K, Brown J, Gerstner E, Rosen B, Kalpathy-Cramer J. Sequential neural networks for biologically-informed glioma segmentation. In: Angelini ED, Landman BA, eds. *Medical Imaging 2018: Image Processing*. Vol 10574. SPIE; 2018:108. doi:10.1117/12.2293941
9. DeepNeuro. https://github.com/QTIM-Lab/DeepNeuro.
10. Beers A, Brown J, Chang K, et al. DeepNeuro: an open-source deep learning toolbox for neuroimaging. August 2018. http://arxiv.org/abs/1808.04589. Accessed August 30, 2018.
11. Chang K, Beers AL, Bai HX, et al. Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bidimensional measurement. *Neuro Oncol*. 2019. doi:10.1093/neuonc/noz106
12. Chang K, Bai HX, Zhou H, et al. Residual Convolutional Neural Network for the Determination of *IDH* Status in Low- and High-Grade Gliomas from MR Imaging. *Clin Cancer Res*. 2018;24(5):1073-1081. doi:10.1158/1078-0432.CCR-17-2236
13. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet (London, England)*. 2006;367(9524):1747-1757. doi:10.1016/S0140-6736(06)68770-9
14. Lloyd-Jones D, Adams R, Carnethon M, et al. Heart Disease and Stroke Statistics--2009 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*. 2009;119(3):480-486. doi:10.1161/CIRCULATIONAHA.108.191259
15. LeCun YA, Bengio Y, Hinton GE. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539
16. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*. 2015;115(3):211-252. doi:10.1007/s11263-015-0816-y

17. Li R, Xing L, Napel S, Rubin D (Daniel L. *Radiomics and Radiogenomics : Technical Basis and Clinical Applications*.
18. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019. doi:10.1038/s41591-018-0300-7
19. Sperr E. PubMed by Year. http://esperr.github.io/pubmed-by-year/. Published 2016.
20. Smith-Bindman R, Kwan ML, Marlow EC, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA - J Am Med Assoc*. 2019. doi:10.1001/jama.2019.11456
21. Reinertsen E, Clifford GD. A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses. *Physiol Meas*. 2018. doi:10.1088/1361-6579/aabf64
22. Reinertsen E, Osipov M, Liu C, Kane JM, Petrides G, Clifford GD. Continuous assessment of schizophrenia using heart rate and accelerometer data. *Physiol Meas*. 2017. doi:10.1088/1361-6579/aa724d
23. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-)*. 2016. doi:10.1126/science.aad0501
24. Webb RC, Ma Y, Krishnan S, et al. Epidermal devices for noninvasive, precise, and continuous mapping of macrovascular and microvascular blood flow. *Sci Adv*. 2015;1(9):e1500701-e1500701. doi:10.1126/sciadv.1500701
25. Zech J, Pain M, Titano J, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. *Radiology*. 2018. doi:10.1148/radiol.2018171093
26. Grossmann P, Stringfield O, El-Hachem N, et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife*. 2017. doi:10.7554/eLife.23421
27. Peter Campbell J, Ataer-Cansizoglu E, Bolon-Canedo V, et al. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. *JAMA Ophthalmol*. 2016. doi:10.1001/jamaophthalmol.2016.0611
28. Chang K, Yoon S, Sheth N, et al. Rapid vs. delayed infrared responses after ischemia reveal recruitment of different vascular beds. *Quant Infrared Thermogr J*. June 2015:1-11. doi:10.1080/17686733.2015.1046677
29. Siless V, Chang K, Fischl B, Yendiki A. AnatomiCuts: Hierarchical clustering of tractography streamlines based on anatomical similarity. *Neuroimage*. 2018;166:32-45. doi:10.1016/j.neuroimage.2017.10.058
30. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*. 1999. doi:10.1006/nimg.1998.0395
31. Liu Y, Syed Z, Scirica BM, Morrow DA, Guttag J V., Stultz CM. ECG morphological variability in beat space for risk stratification after acute coronary syndrome. *J Am Heart Assoc*. 2014. doi:10.1161/JAHA.114.000981
32. Webber WRS, Litt B, Lesser RP, Fisher RS, Bankman I. Automatic EEG spike detection: what should the computer imitate? *Electroencephalogr Clin Neurophysiol*. 1993. doi:10.1016/0013-4694(93)90149-P
33. Liu C, Oster J, Reinertsen E, et al. A comparison of entropy approaches for AF discrimination. *Physiol Meas*. 2018. doi:10.1088/1361-6579/aacc48
34. Poplin R, Varadarajan A V., Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018. doi:10.1038/s41551-018-0195-0
35. Mitani A, Huang A, Venugopalan S, et al. Detection of anaemia from retinal fundus

images via deep learning. *Nat Biomed Eng*. December 2019. doi:10.1038/s41551-019-0487-z

36. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. Sheikh A, ed. *PLOS Med*. 2018;15(11):e1002686. doi:10.1371/journal.pmed.1002686
37. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama*. 2016;304(6):649-656. doi:10.1001/jama.2016.17216
38. Li MD, Chang K, Bearce B, et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *npj Digit Med*. 2020;3(1):48. doi:10.1038/s41746-020-0255-1
39. Zhao Y, Chang M, Wang R, et al. Deep Learning Based on MRI for Differentiation of Low- and High-Grade in Low-Stage Renal Cell Carcinoma. *J Magn Reson Imaging*. March 2020. doi:10.1002/jmri.27153
40. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056
41. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci*. 2018;115(13):E2970-E2979. doi:10.1073/pnas.1717139115
42. Chang K, Beers AL, Bai HX, et al. Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bi-dimensional measurement. *Neuro Oncol*. June 2019. doi:10.1093/neuonc/noz106
43. Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digit Med*. 2019. doi:10.1038/s41746-019-0172-3
44. Lu MT, Ivanov A, Mayrhofer T, Hosny A, Aerts HJWL, Hoffmann U. Deep Learning to Assess Long-term Mortality From Chest Radiographs. *JAMA Netw Open*. 2019. doi:10.1001/jamanetworkopen.2019.7416
45. Brown JM, Campbell JP, Beers A, et al. Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning. In: *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*. Vol 10579. ; 2018. doi:10.1117/12.2295942
46. Brown JM, Kalpathy-Cramer J, Campbell JP, et al. Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning. In: Zhang J, Chen P-H, eds. *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. Vol 10579. SPIE; 2018:22. doi:10.1117/12.2295942
47. Xi IL, Zhao Y, Wang R, et al. Deep learning to distinguish benign from malignant renal lesions based on routine MR imaging. *Clin Cancer Res*. January 2020:clincanres.0374.2019. doi:10.1158/1078-0432.CCR-19-0374
48. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. *Nature*. 2018;555(7697):487-492. doi:10.1038/nature25988
49. Coyner AS, Swan R, Campbell JP, et al. Automated Fundus Image Quality Assessment in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *Ophthalmol Retin*. 2019;3(5):444-450. doi:10.1016/j.oret.2019.01.015
50. Morales MA, Izquierdo-Garcia D, Aganj I, Kalpathy-Cramer J, Rosen BR, Catana C. Implementation and Validation of a Three-dimensional Cardiac Motion Estimation

Network. *Radiol Artif Intell*. 2019. doi:10.1148/ryai.2019180080

51. Cowan IA, MacDonald SLS, Floyd RA. Measuring and managing radiologist workload: Measuring radiologist reporting times using data from a Radiology Information System. *J Med Imaging Radiat Oncol*. 2013. doi:10.1111/1754-9485.12092
52. Rosenkrantz AB, Duszak R, Babb JS, Glover M, Kang SK. Discrepancy Rates and Clinical Impact of Imaging Secondary Interpretations: A Systematic Review and Meta-Analysis. *J Am Coll Radiol*. 2018;15(9):1222-1231. doi:10.1016/j.jacr.2018.05.037
53. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al. Plus Disease in Retinopathy of Prematurity: A Continuous Spectrum of Vascular Abnormality as a Basis of Diagnostic Variability. *Ophthalmology*. 2016;123(11):2338-2344. doi:10.1016/j.ophtha.2016.07.026
54. Wright RW, Ross JR, Haas AK, et al. Osteoarthritis classification scales: Interobserver reliability and arthroscopic correlation. *J Bone Jt Surg - Am Vol*. 2014. doi:10.2106/JBJS.M.00929
55. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi:10.1038/ncomms5006
56. Silva MA, Patel J, Kavouridis V, et al. Machine Learning Models can Detect Aneurysm Rupture and Identify Clinical Features Associated with Rupture. *World Neurosurg*. 2019. doi:10.1016/j.wneu.2019.06.231
57. Bakas S, Zeng K, Sotiras A, et al. GLISTRboost: Combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 9556. Springer Verlag; 2016:144-155. doi:10.1007/978-3-319-30858-6_13
58. Zhou H, Vallières M, Bai HX, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro Oncol*. 2017;19(6):862-870. doi:10.1093/neuonc/now256
59. Coroller TP, Grossmann P, Hou Y, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol*. 2015. doi:10.1016/j.radonc.2015.02.015
60. Chang K, Zhang B, Guo X, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. *Neuro Oncol*. 2016;18(12):1680-1687. doi:10.1093/neuonc/now086
61. Kalpathy-Cramer J, Mamomov A, Zhao B, et al. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. *Tomogr a J imaging Res*. 2016;2(4):430-437. doi:10.18383/j.tom.2016.00235
62. Emaminejad N, Kim GH, Brown MS, McNitt-Gray MF, Hoffman J, Wahi-Anwar M. The effects of variations in parameters and algorithm choices on calculated radiomics feature values: initial investigations and comparisons to feature variability across CT image acquisition conditions. In: Mori K, Petrick N, eds. *Medical Imaging 2018: Computer-Aided Diagnosis*. SPIE; 2018:140. doi:10.1117/12.2293864
63. Chang K, Beers A, Brown J, Kalpathy-Cramer J. Resources and datasets for radiomics. In: *Radiomics and Radiogenomics*. Chapman and Hall/CRC; 2019:179-189. doi:10.1201/9781351208277-11
64. Kim H, Park CM, Lee M, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: Analysis of intra- and inter-reader variability and inter-

reconstruction algorithm variability. *PLoS One*. 2016;11(10). doi:10.1371/journal.pone.0164924

65. Mackin D, Fave X, Zhang L, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. *Invest Radiol*. 2015;50(11):757-765. doi:10.1097/RLI.0000000000000180
66. Zhao B, Tan Y, Tsai W-Y, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6(1):23428. doi:10.1038/srep23428
67. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*. April 2018:172361. doi:10.1148/radiol.2018172361
68. Shafiq-ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;44(3):1050-1062. doi:10.1002/mp.12123
69. Fave X, Mackin D, Yang J, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? *Med Phys*. 2015;42(12):6784-6797. doi:10.1118/1.4934826
70. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother Oncol*. 2019;130:2-9. doi:10.1016/j.radonc.2018.10.027
71. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology*. 2019. doi:10.1148/radiol.2018181352
72. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. 2017;318(22):2199. doi:10.1001/jama.2017.14585
73. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024. doi:10.1109/TMI.2014.2377694
74. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. January 2019. http://arxiv.org/abs/1901.07031. Accessed October 30, 2019.
75. Gurovich Y, Hanani Y, Bar O, et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med*. 2019. doi:10.1038/s41591-018-0279-0
76. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun*. 2019. doi:10.1038/s41467-019-10933-3
77. Schwarz CG, Kremers WK, Therneau TM, et al. Identification of Anonymous MRI Research Participants with Face-Recognition Software. *N Engl J Med*. 2019;381(17):1684-1686. doi:10.1056/NEJMc1908881
78. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med*. 2018. doi:10.1371/journal.pmed.1002683
79. Albadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing: Impact. *Med Phys*. 2018. doi:10.1002/mp.12752

80. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. *Korean J Radiol*. 2019;20(3):405-410. doi:10.3348/kjr.2019.0025
81. De Robles P, Fiest KM, Frolkis AD, et al. The worldwide incidence and prevalence of primary brain tumors: A systematic review and meta-analysis. *Neuro Oncol*. 2015;17(6):776-783. doi:10.1093/neuonc/nou283
82. Thakkar JP, Dolecek TA, Horbinski C, et al. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol Biomarkers Prev*. 2014;23(10):1985-1996. doi:10.1158/1055-9965.EPI-14-0275
83. Cancer Genome Atlas Research Network, Brat DJ, Verhaak RGW, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med*. 2015;372(26):2481-2498. doi:10.1056/NEJMoa1402121
84. Omuro A, CA D, WK Y, AJ G, AC W, CA A. Glioblastoma and Other Malignant Gliomas. *JAMA*. 2013;310(17):1842. doi:10.1001/jama.2013.280319
85. Johnson DR, O'Neill BP. Glioblastoma survival in the United States before and during the temozolomide era. *J Neurooncol*. 2012;107(2):359-364. doi:10.1007/s11060-011-0749-4
86. Ostrom QT, Gittleman H, Fulop J, et al. CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. *Neuro Oncol*. 2015;17 Suppl 4(suppl 4):iv1-iv62. doi:10.1093/neuonc/nov189
87. Alexander BM, Cloughesy TF. Adult Glioblastoma. *J Clin Oncol*. 2017;35(21):2402-2409. doi:10.1200/JCO.2017.73.0119
88. Ellingson BM, Wen PY, Cloughesy TF. Modified Criteria for Radiographic Response Assessment in Glioblastoma Clinical Trials. *Neurotherapeutics*. 2017;14(2):307-320. doi:10.1007/s13311-016-0507-6
89. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol*. 2010;28(11):1963-1972. doi:10.1200/JCO.2009.26.3541
90. Dempsey MF, Condon BR, Hadley DM. Measurement of tumor &quot;size&quot; in recurrent malignant glioma: 1D, 2D, or 3D? *AJNR Am J Neuroradiol*. 2005;26(4):770-776. http://www.ncbi.nlm.nih.gov/pubmed/15814919. Accessed January 2, 2018.
91. Provenzale JM, Ison C, DeLong D. Bidimensional Measurements in Brain Tumors: Assessment of Interobserver Variability. *Am J Roentgenol*. 2009;193(6):W515-W522. doi:10.2214/AJR.09.2615
92. Provenzale JM, Mancini MC. Assessment of intra-observer variability in measurement of high-grade brain tumors. *J Neurooncol*. 2012;108(3):477-483. doi:10.1007/s11060-012-0843-2
93. Reuter M, Gerstner ER, Rapalino O, Batchelor TT, Rosen B, Fischl B. Impact of MRI head placement on glioma response assessment. *J Neurooncol*. 2014;118(1):123-129. doi:10.1007/s11060-014-1403-8
94. Deeley MA, Chen A, Datteri R, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. *Phys Med Biol*. 2011;56(14):4557-4577. doi:10.1088/0031-9155/56/14/021
95. Huang RY, Rahman R, Ballman K V., et al. The Impact of T2/FLAIR Evaluation per RANO Criteria on Response Assessment of Recurrent Glioblastoma Patients Treated with Bevacizumab. *Clin Cancer Res*. 2016;22(3):575-581. doi:10.1158/1078-0432.CCR-14-

3040

96. Bakas S, Reyes M, Jakab A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. November 2018. http://arxiv.org/abs/1811.02629. Accessed December 15, 2019.

97. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024. doi:10.1109/TMI.2014.2377694

98. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61-78. doi:10.1016/j.media.2016.10.004

99. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with Deep Neural Networks. *Med Image Anal*. 2017;35:18-31. doi:10.1016/j.media.2016.05.004

100. Parsons DW, Jones S, Zhang X, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science (80-)*. 2008;321(5897):1807-1812. doi:10.1126/science.1164382

101. Eckel-Passow JE, Lachance DH, Molinaro AM, et al. Glioma Groups Based on 1p/19q, *IDH* , and *TERT* Promoter Mutations in Tumors. *N Engl J Med*. 2015;372(26):2499-2508. doi:10.1056/NEJMoa1407279

102. Yang H, Ye D, Guan K-L, Xiong Y. IDH1 and IDH2 Mutations in Tumorigenesis: Mechanistic Insights and Clinical Perspectives. *Clin Cancer Res*. 2012;18(20):5562-5571. doi:10.1158/1078-0432.CCR-12-1773

103. Flavahan WA, Drier Y, Liau BB, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*. 2016;529(7584):110-114. doi:10.1038/nature16490

104. Hartmann C, Hentschel B, Wick W, et al. Patients with IDH1 wild type anaplastic astrocytomas exhibit worse prognosis than IDH1-mutated glioblastomas, and IDH1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas. *Acta Neuropathol*. 2010;120(6):707-718. doi:10.1007/s00401-010-0781-z

105. Houillier C, Wang X, Kaloshi G, et al. IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. *Neurology*. 2010;75(17):1560-1566. doi:10.1212/WNL.0b013e3181f96282

106. Yan H, Parsons DW, Jin G, et al. *IDH1* and *IDH2* Mutations in Gliomas. *N Engl J Med*. 2009;360(8):765-773. doi:10.1056/NEJMoa0808710

107. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol*. 2016;131(6):803-820. doi:10.1007/s00401-016-1545-1

108. SongTao Q, Lei Y, Si G, et al. IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer Sci*. 2012;103(2):269-273. doi:10.1111/j.1349-7006.2011.02134.x

109. Narita Y, Narita Y, Miyakita Y, et al. IDH1/2 mutation is a prognostic marker for survival and predicts response to chemotherapy for grade II gliomas concomitantly treated with radiation therapy. *Int J Oncol*. 2012;41(4):1325-1336. doi:10.3892/ijo.2012.1564

110. Molenaar RJ, Botman D, Smits MA, et al. Radioprotection of IDH1-Mutated Cancer Cells by the IDH1-Mutant Inhibitor AGI-5198. *Cancer Res*. 2015;75(22):4790-4802. doi:10.1158/0008-5472.CAN-14-3603

111. Mohrenz IV, Antonietti P, Pusch S, et al. Isocitrate dehydrogenase 1 mutant R132H sensitizes glioma cells to BCNU-induced oxidative stress and cell death. *Apoptosis*. 2013;18(11):1416-1425. doi:10.1007/s10495-013-0877-8

112. Sulkowski PL, Corso CD, Robinson ND, et al. 2-Hydroxyglutarate produced by neomorphic IDH mutations suppresses homologous recombination and induces PARP inhibitor sensitivity. *Sci Transl Med*. 2017;9(375). http://stm.sciencemag.org/content/9/375/eaal2463. Accessed July 16, 2017.

113. Beiko J, Suki D, Hess KR, et al. IDH1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection. *Neuro Oncol*. 2014;16(1):81-91. doi:10.1093/neuonc/not159

114. Heiss W, Kidwell CS. Imaging for prediction of functional outcome and assessment of recovery in ischemic stroke. *Stroke*. 2014;45(4):1195-1201. doi:10.1161/STROKEAHA.113.003611

115. Poisson SN, Johnston SC, Josephson SA. Urinary tract infections complicating stroke: Mechanisms, consequences, and possible solutions. *Stroke*. 2010;41(4). doi:10.1161/STROKEAHA.109.576413

116. Lansberg MG, Albers GW, Beaulieu C, Marks MP. Comparison of diffusion-weighted MRI and CT in acute stroke. *Neurology*. 2000;54(8):1557-1561. doi:10.1212/WNL.54.8.1557

117. Demaerschalk BM, Cheng NT, Kim AS. Intravenous Thrombolysis for Acute Ischemic Stroke Within 3 Hours Versus Between 3 and 4.5 Hours of Symptom Onset. *The Neurohospitalist*. 2015;5(3):101-109. doi:10.1177/1941874415583116

118. Martel AL, Allder SJ, Delay GS, Morgan PS, Moody AR. *Measurement of Infarct Volume in Stroke Patients Using Adaptive Segmentation of Diffusion Weighted MR Images*. Vol 1679.; 1999.

119. Jacobs MA, Knight RA, Soltanian-Zadeh H, et al. Unsupervised segmentation of multiparameter MRI in experimental cerebral ischemia with comparison to T2, diffusion, and ADC MRI parameters and histopathological validation. *J Magn Reson Imaging*. 2000;11(4):425-437. http://www.ncbi.nlm.nih.gov/pubmed/10767072. Accessed June 7, 2018.

120. Maier O, Menze BH, von der Gablentz J, et al. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med Image Anal*. 2017;35:250-269. doi:10.1016/j.media.2016.07.009

121. Chen L, Bentley P, Rueckert D. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *NeuroImage Clin*. 2017;15:633-643. doi:10.1016/j.nicl.2017.06.016

122. Lindsberg PJ, Roine RO. Hyperglycemia in Acute Stroke. *Stroke*. 2004;35(2):363-364. doi:10.1161/01.STR.0000115297.92132.84

123. Su Y-C, Huang K-F, Yang F-Y, Lin S-K. Elevation of troponin I in acute ischemic stroke. *PeerJ*. 2016;4:e1866. doi:10.7717/peerj.1866

124. Chamorro Á, Urra X, Planas AM. Infection {After} {Acute} {Ischemic} {Stroke}. *Stroke*. 2007;38(3):1097-1103. doi:10.1161/01.STR.0000258346.68966.9d

125. Bogason E, Morrison K, Zalatimo O, et al. Urinary Tract Infections in Hospitalized Ischemic Stroke Patients: Source and Impact on Outcome. *Cureus*. 2017;9(2):e1014. doi:10.7759/cureus.1014

126. Sykora M, Siarnik P, Diedler J, et al. -Blockers, Pneumonia, and Outcome After Ischemic

Stroke: Evidence From Virtual International Stroke Trials Archive. *Stroke*. 2015;46(5):1269-1274. doi:10.1161/STROKEAHA.114.008260

127. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016;316(22):2402. doi:10.1001/jama.2016.17216

128. Miotto R, Li L, Kidd BA, Dudley JT, Agarwal P. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep*. 2016;6(1):26094. doi:10.1038/srep26094

129. Dluhoš P, Schwarz D, Cahn W, et al. Multi-center machine learning in imaging psychiatry: A meta-model approach. *Neuroimage*. 2017;155:10-24. doi:10.1016/j.neuroimage.2017.03.027

130. Xia W, Wan Z, Yin Z, et al. It's all in the timing: calibrating temporal penalties for biomedical data sharing. *J Am Med Informatics Assoc*. 2018;25(1):25-31. doi:10.1093/jamia/ocx101

131. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 11383 LNCS. Springer Verlag; 2019:92-104. doi:10.1007/978-3-030-11723-8_9

132. Hsieh K, Phanishayee A, Mutlu O, Gibbons PB. The Non-IID Data Quagmire of Decentralized Machine Learning. September 2019. http://arxiv.org/abs/1910.00189. Accessed December 27, 2019.

133. Kairouz P, McMahan HB, Avent B, et al. Advances and Open Problems in Federated Learning. December 2019. http://arxiv.org/abs/1912.04977. Accessed December 24, 2019.

134. Kaggle. Diabetic Retinopathy Detection. https://www.kaggle.com/c/diabetic-retinopathy-detection. Published 2015.

135. Quellec G, Charrière K, Boudi Y, Cochener B, Lamard M. Deep image mining for diabetic retinopathy screening. *Med Image Anal*. 2017;39:178-193. doi:10.1016/j.media.2017.04.012

136. Graham B. *Kaggle Diabetic Retinopathy Detection Competition Report*.; 2015.

137. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2016:770-778. doi:10.1109/CVPR.2016.90

138. Chollet F. Keras: Deep Learning library for Theano and TensorFlow. *GitHub Repos*. 2015:1-21.

139. Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*. 2016:19. http://arxiv.org/abs/1605.02688.

140. Glorot X, Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proc Int Conf Artif Intell Stat (AISTATS'10) Soc Artif Intell Stat*. 2010. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.207.2059. Accessed April 12, 2017.

141. Dietterich TG. Ensemble methods in machine learning. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* ; 2000. doi:10.1007/3-540-45014-9_1

142. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics). ; 2018. doi:10.1007/978-3-030-01424-7_27

143. USF Digital Mammography. DDSM: Digital Database for Screening Mammography. http://marathon.csee.usf.edu/Mammography/Database.html.
144. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 07-12-June. ; 2015:1-9. doi:10.1109/CVPR.2015.7298594
145. Kingma DP, Ba JL. Adam: a Method for Stochastic Optimization. *Int Conf Learn Represent 2015*. 2015:1-15. doi:http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503
146. Ratcliff R. Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychol Rev*. 1990. doi:10.1037/0033-295X.97.2.285
147. Hansen LK, Salamon P. Neural Network Ensembles. *IEEE Trans Pattern Anal Mach Intell*. 1990;12(10):993-1001. doi:10.1109/34.58871
148. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191
149. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Cha KH, Richter CD. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Phys Med Biol*. 2017;62(23):8894-8908. doi:10.1088/1361-6560/aa93d4
150. Baldi P. Autoencoders, Unsupervised Learning, and Deep Architectures. *ICML Unsupervised Transf Learn*. 2012. doi:10.1561/2200000006
151. Lecouat B, Chang K, Foo C-S, et al. Semi-Supervised Deep Learning for Abnormality Classification in Retinal Images. December 2018. http://arxiv.org/abs/1812.07832. Accessed December 15, 2019.
152. Sutskever I. Training Recurrent neural Networks. *PhD thesis*. 2013.
153. Shokri R, Shmatikov V. Privacy-preserving deep learning. In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015*. ; 2016. doi:10.1109/ALLERTON.2015.7447103
154. Brendan McMahan H, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*. ; 2017.
155. Chen J, Pan X, Monga R, Bengio S, Jozefowicz R. Revisiting Distributed Synchronous SGD. April 2016. http://arxiv.org/abs/1604.00981. Accessed December 24, 2019.
156. Dean J, Corrado GS, Monga R, et al. Large Scale Distributed Deep Networks. *NIPS 2012 Neural Inf Process Syst*. 2012:1-11. doi:10.1109/ICDAR.2011.95
157. Gupta O, Raskar R. Distributed learning of deep neural network over multiple agents. *J Netw Comput Appl*. 2018. doi:10.1016/j.jnca.2018.05.003
158. Vepakomma P, Gupta O, Swedish T, Raskar R. Split learning for health: Distributed deep learning without sharing raw patient data. December 2018. http://arxiv.org/abs/1812.00564. Accessed July 22, 2019.
159. Vepakomma P, Swedish T, Raskar R, Gupta O, Dubey A. No Peek: A Survey of private distributed deep learning. December 2018. http://arxiv.org/abs/1812.03288. Accessed January 1, 2020.
160. Singh A, Vepakomma P, Gupta O, Raskar R. Detailed comparison of communication

efficiency of split learning and federated learning. September 2019. http://arxiv.org/abs/1909.09145. Accessed January 3, 2020.

161. Poirot MG, Vepakomma P, Chang K, Kalpathy-Cramer J, Gupta R, Raskar R. Split Learning for collaborative deep learning in healthcare. December 2019. http://arxiv.org/abs/1912.12115. Accessed December 31, 2019.

162. Hang Su HC. Experiments on Parallel Training of Deep Neural Network Using Model Averaging. *ArXiv*. 2015:1-6. http://arxiv.org/abs/1507.01239.

163. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. February 2016. http://arxiv.org/abs/1602.07261. Accessed August 12, 2018.

164. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. http://proceedings.mlr.press/v37/ioffe15.pdf. Accessed April 12, 2017.

165. Bjorck J, Gomes C, Selman B, Weinberger KQ. Understanding batch normalization. In: *Advances in Neural Information Processing Systems*. ; 2018.

166. Santurkar S, Tsipras D, Ilyas A, Madry A. How does batch normalization help optimization? In: *Advances in Neural Information Processing Systems*. ; 2018.

167. Wu Y, He K. Group normalization. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ; 2018. doi:10.1007/978-3-030-01261-8_1

168. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: *30th International Conference on Machine Learning, ICML 2013*. ; 2013.

169. Yu H, Jin R, Yang S. On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization. May 2019. http://arxiv.org/abs/1905.03817. Accessed December 30, 2019.

170. Goodfellow IJ, Mirza M, Xiao D, Courville A, Bengio Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In: *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. ; 2014.

171. Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci U S A*. 2017. doi:10.1073/pnas.1611835114

172. Zeng G, Chen Y, Cui B, Yu S. Continual learning of context-dependent processing in neural networks. *Nat Mach Intell*. 2019. doi:10.1038/s42256-019-0080-x

173. Mallya A, Davis D, Lazebnik S. Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ; 2018. doi:10.1007/978-3-030-01225-0_5

174. Karani N, Chaitanya K, Baumgartner C, Konukoglu E. A lifelong learning approach to brain MR segmentation across scanners and protocols. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ; 2018. doi:10.1007/978-3-030-00928-1_54

175. Kamnitsas K, Baumgartner C, Ledig C, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ; 2017. doi:10.1007/978-3-319-59050-9_47

176. Zhao H, Combes RT des, Zhang K, Gordon GJ. On Learning Invariant Representation for Domain Adaptation. January 2019. http://arxiv.org/abs/1901.09453. Accessed January 21, 2020.

177. Sharma V, Vepakomma P, Swedish T, Chang K, Kalpathy-Cramer J, Raskar R. ExpertMatcher: Automating ML Model Selection for Users in Resource Constrained Countries. October 2019. http://arxiv.org/abs/1910.02312. Accessed February 9, 2020.

178. Sharma V, Vepakomma P, Swedish T, Chang K, Kalpathy-Cramer J, Raskar R. ExpertMatcher: Automating ML Model Selection for Clients using Hidden Representations. October 2019. http://arxiv.org/abs/1910.03731. Accessed February 9, 2020.

179. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*. 2002. doi:10.1613/jair.953

180. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. MixUp: Beyond empirical risk minimization. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. ; 2018.

181. Chang K, Bai HX, Zhou H, et al. Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from mr imaging. *Clin Cancer Res*. 2018;24(5):1073-1081. doi:10.1158/1078-0432.CCR-17-2236

182. Beers A, Brown J, Chang K, et al. High-resolution medical image synthesis using progressively grown generative adversarial networks. May 2018. http://arxiv.org/abs/1805.03144. Accessed May 23, 2018.

183. Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng*. 2019. doi:10.1038/s41551-018-0324-9

184. Shan H, Padole A, Homayounieh F, et al. Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. *Nat Mach Intell*. 2019. doi:10.1038/s42256-019-0057-9

185. Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci Rep*. 2019. doi:10.1038/s41598-019-52737-x

186. Zhang Y, Wu H, Liu H, Tong L, Wang MD. Improve Model Generalization and Robustness to Dataset Bias with Bias-regularized Learning and Domain-guided Augmentation. October 2019. http://arxiv.org/abs/1910.06745. Accessed December 31, 2019.

187. Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: *Proceedings of the ACM Conference on Computer and Communications Security*. ; 2017. doi:10.1145/3133956.3134077

188. Zhu L, Liu Z, Han S. Deep Leakage from Gradients. June 2019. http://arxiv.org/abs/1906.08935. Accessed January 1, 2020.

189. Vepakomma P, Gupta O, Dubey A, Raskar R. Reducing leakage in distributed deep learning for sensitive health data. In: *ICLR AI for Social Good Workshop 2019*. ; 2019.

190. Wood A, Altman M, Bembenek A, et al. Differential Privacy: A Primer for a Non-Technical Audience. *SSRN Electron J*. 2019. doi:10.2139/ssrn.3338027

191. Abadi M, McMahan HB, Chu A, et al. Deep learning with differential privacy. In: *Proceedings of the ACM Conference on Computer and Communications Security*. ; 2016. doi:10.1145/2976749.2978318

192. Wu B, Zhao S, Sun G, et al. P3SGD: Patient Privacy Preserving SGD for Regularizing Deep CNNs in Pathological Image Classification. May 2019. http://arxiv.org/abs/1905.12883. Accessed January 1, 2020.
193. Beaulieu-Jones BK, Yuan W, Finlayson SG, Wu ZS. Privacy-Preserving Distributed Deep Learning for Clinical Data. December 2018. http://arxiv.org/abs/1812.01484. Accessed January 1, 2020.
194. Beaulieu-Jones BK, Wu ZS, Williams C, et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circ Cardiovasc Qual Outcomes*. 2019;12(7):e005122. doi:10.1161/CIRCOUTCOMES.118.005122
195. Badawi A Al, Chao J, Lin J, et al. The AlexNet Moment for Homomorphic Encryption: HCNN, the First Homomorphic CNN on Encrypted Data with GPUs. November 2018. http://arxiv.org/abs/1811.00778. Accessed January 1, 2020.
196. Chao J, Badawi A Al, Unnikrishnan B, et al. CaRENets: Compact and Resource-Efficient CNN for Homomorphic Inference on Encrypted Medical Images. January 2019. http://arxiv.org/abs/1901.10074. Accessed January 1, 2020.
197. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(1):7-34. doi:10.3322/caac.21551
198. Duffy SW, Tabár L, Chen H-H, et al. The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties. *Cancer*. 2002;95(3):458-469. doi:10.1002/cncr.10765
199. Tabár L, Vitak B, Chen HH, Yen MF, Duffy SW, Smith RA. Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer*. 2001;91(9):1724-1731. http://www.ncbi.nlm.nih.gov/pubmed/11335897. Accessed May 20, 2019.
200. Boyd NF, Byng JW, Jong RA, et al. Quantitative Classification of Mammographic Densities and Breast Cancer Risk: Results From the Canadian National Breast Screening Study. *JNCI J Natl Cancer Inst*. 1995;87(9):670-675. doi:10.1093/jnci/87.9.670
201. Razzaghi H, Troester MA, Gierach GL, Olshan AF, Yankaskas BC, Millikan RC. Mammographic density and breast cancer risk in White and African American Women. *Breast Cancer Res Treat*. 2012. doi:10.1007/s10549-012-2185-3
202. Lehman CD, Yala A, Schuster T, et al. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology*. 2019;290(1):52-58. doi:10.1148/radiol.2018180694
203. Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). *Radiol Clin North Am*. 2002. doi:10.1016/S0033-8389(01)00017-3
204. Sprague BL, Conant EF, Onega T, et al. Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice. *Ann Intern Med*. 2016;165(7):457. doi:10.7326/M15-2934
205. Brandt KR, Scott CG, Ma L, et al. Comparison of Clinical and Automated Breast Density Measurements: Implications for Risk Prediction and Supplemental Screening. *Radiology*. 2016;279(3):710-719. doi:10.1148/radiol.2015151261
206. Youk JH, Gweon HM, Son EJ, Kim J-A. Automated Volumetric Breast Density Measurements in the Era of the BI-RADS Fifth Edition: A Comparison With Visual Assessment. *Am J Roentgenol*. 2016;206(5):1056-1062. doi:10.2214/AJR.15.15472
207. Brown JM, Campbell JP, Beers A, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA*

Ophthalmol. May 2018. doi:10.1001/jamaophthalmol.2018.1934

208. Langlotz CP, Allen B, Erickson BJ, et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology*. 2019;291(3):781-791. doi:10.1148/radiol.2019190613
209. Allen B, Agarwal S, Kalpathy-Cramer J, Dreyer K. Democratizing AI. *J Am Coll Radiol*. 2019;16(7):961-963. doi:10.1016/j.jacr.2019.04.023
210. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *N Engl J Med*. 2005;353(17):1773-1783. doi:10.1056/NEJMoa052911
211. Abadi M, Barham P, Chen J, et al. TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning. *Proc 12th USENIX Conf Oper Syst Des Implement*. 2016.
212. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely Connected Convolutional Networks. August 2016. http://arxiv.org/abs/1608.06993. Accessed March 1, 2017.
213. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. ; 2016. doi:10.1109/CVPR.2016.308
214. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Int Conf Learn Represent*. 2015:1-14. doi:10.1016/j.infsof.2008.09.005
215. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*. 2015. doi:10.1007/s11263-015-0816-y
216. Cheng J, Wang Z, Pollastri G. A neural network approach to ordinal regression. In: *Proceedings of the International Joint Conference on Neural Networks*. ; 2008. doi:10.1109/IJCNN.2008.4633963
217. McInnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. February 2018. http://arxiv.org/abs/1802.03426. Accessed September 23, 2018.
218. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977. doi:10.2307/2529310
219. Del Carmen MG, Halpern EF, Kopans DB, et al. Mammographic breast density and race. *Am J Roentgenol*. 2007. doi:10.2214/AJR.06.0619
220. Jia Deng, Wei Dong, Socher R, Li-Jia Li, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ; 2009. doi:10.1109/CVPRW.2009.5206848
221. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. *Radiology*. 2019. doi:10.1148/radiol.2018181422
222. Hu S-Y, Beers A, Chang K, et al. Deep feature transfer between localization and segmentation tasks. November 2018. http://arxiv.org/abs/1811.02539. Accessed December 24, 2019.
223. Brown JM, Campbell JP, Beers A, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA Ophthalmol*. 2018;97239:1-8. doi:10.1001/jamaophthalmol.2018.1934
224. Van Hulse J, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. In: ; 2008. doi:10.1145/1273496.1273614

225. Wu N, Phang J, Park J, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. March 2019. http://arxiv.org/abs/1903.08297. Accessed June 11, 2019.

226. Mohamed AA, Berg WA, Peng H, Luo Y, Jankowitz RC, Wu S. A deep learning method for classifying mammographic breast density categories. *Med Phys*. 2018. doi:10.1002/mp.12683

227. Keavey E, Phelan N, O'Connell AM, et al. Comparison of the clinical performance of three digital mammography systems in a breast cancer screening programme. *Br J Radiol*. 2012. doi:10.1259/bjr/29747759

228. Kim JS, Greene MJ, Zlateski A, et al. Space-time wiring specificity supports direction selectivity in the retina. *Nature*. 2014. doi:10.1038/nature13240

229. Norman TC, Bountra C, Edwards AM, Yamamoto KR, Friend SH. Leveraging crowdsourcing to facilitate the discovery of new medicines. In: *Science Translational Medicine*. ; 2011. doi:10.1126/scitranslmed.3002678

230. Farahani K, Kalpathy-Cramer J, Chenevert TL, et al. Computational Challenges and Collaborative Projects in the NCI Quantitative Imaging Network. *Tomogr (Ann Arbor, Mich)*. 2016;2(4):242-249. doi:10.18383/j.tom.2016.00265

231. Jin H, Song Q, Hu X. Auto-keras: An efficient neural architecture search system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ; 2019. doi:10.1145/3292500.3330648

232. He K, Girshick R, Dollár P. Rethinking ImageNet Pre-training. November 2018. http://arxiv.org/abs/1811.08883. Accessed June 9, 2019.

233. Kamnitsas K, Bai W, Ferrante E, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ; 2018. doi:10.1007/978-3-319-75238-9_38

234. Isensee F, Petersen J, Klein A, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. In: *Informatik Aktuell*. ; 2019. doi:10.1007/978-3-658-25326-4_7

235. Bakker MF, de Lange S V., Pijnappel RM, et al. Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. *N Engl J Med*. 2019;381(22):2091-2102. doi:10.1056/NEJMoa1903986

236. Choudhery S, Chou SHS, Chang K, Kalpathy-Cramer J, Lehman CD. Kinetic Analysis of Lesions Identified on a Rapid Abridged Multiphase (RAMP) Breast MRI Protocol. *Acad Radiol*. 2019. doi:10.1016/j.acra.2019.05.001

237. Wen PY, Macdonald DR, Reardon DA, et al. Updated Response Assessment Criteria for High-Grade Gliomas: Response Assessment in Neuro-Oncology Working Group. *J Clin Oncol*. 2010;28(11):1963-1972. doi:10.1200/JCO.2009.26.3541

238. Vos MJ, Uitdehaag BMJ, Barkhof F, et al. Interobserver variability in the radiological assessment of response to chemotherapy in glioma. *Neurology*. 2003;60(5):826-830. doi:10.1212/01.WNL.0000049467.54667.92

239. Boxerman JL, Zhang Z, Safriel Y, et al. Early post-bevacizumab progression on contrast-enhanced MRI as a prognostic marker for overall survival in recurrent glioblastoma: results from the ACRIN 6677/RTOG 0625 Central Reader Study. *Neuro Oncol*. 2013;15(7):945-954. doi:10.1093/neuonc/not049

240. Barboriak DP, Zhang Z, Desai P, et al. Interreader Variability of Dynamic Contrast-

enhanced MRI of Recurrent Glioblastoma: The Multicenter ACRIN 6677/RTOG 0625 Study. *Radiology*. November 2018:181296. doi:10.1148/radiol.2019181296

241. Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro Oncol*. 2015;17(9):1188-1198. doi:10.1093/neuonc/nov095

242. Zhang B, Chang K, Ramkissoon S, et al. Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro Oncol*. 2017;19(1):109-117. doi:10.1093/neuonc/now121

243. Grossmann P, Narayan V, Chang K, et al. Quantitative Imaging Biomarkers for Risk Stratification of Patients with Recurrent Glioblastoma Treated with Bevacizumab. *Neuro Oncol*. 2017:1-32. doi:10.1093/neuonc/nox092

244. Smits M, van den Bent MJ. Imaging Correlates of Adult Glioma Genotypes. *Radiology*. 2017;284(2):316-331. doi:10.1148/radiol.2017151930

245. Zhou H, Vallières M, Bai HX, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro Oncol*. 2017;19(6):862-870. doi:10.1093/neuonc/now256

246. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with Deep Neural Networks. *Med Image Anal*. 2017;35:18-31. doi:10.1016/j.media.2016.05.004

247. Kleesiek J, Urban G, Hubert A, et al. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *Neuroimage*. 2016;129:460-469. doi:10.1016/j.neuroimage.2016.01.024

248. Jenkinson M, Beckmann CF, Behrens TEJJ, Woolrich MW, Smith SM. FSL. *Neuroimage*. 2012;62(2):782-790. doi:10.1016/j.neuroimage.2011.09.015

249. Ségonne F, Dale AM, Busa E, et al. A hybrid approach to the skull stripping problem in MRI. *Neuroimage*. 2004;22(3):1060-1075. doi:10.1016/j.neuroimage.2004.03.032

250. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res*. 1996;29(3):162-173. doi:10.1006/cbmr.1996.0014

251. Shattuck DW, Leahy RM. Brainsuite: An automated cortical surface identification tool. *Med Image Anal*. 2002;6(2):129-142. doi:10.1016/S1361-8415(02)00054-3

252. Iglesias JE, Liu CY, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging*. 2011;30(9):1617-1634. doi:10.1109/TMI.2011.2138152

253. Learning UD. Automated Segmentation of Hyperintense Regions in FLAIR MRI Using Deep Learning. *Tomogr a J imaging Res*. 2016;2(4):334-340. doi:10.18383/j.tom.2016.00166

254. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7

255. Ina Ly K, Vakulenko-Lagun B, Emblem KE, et al. Probing tumor microenvironment in patients with newly diagnosed glioblastoma during chemoradiation and adjuvant temozolomide with functional MRI. *Sci Rep*. 2018;8(1):17062. doi:10.1038/s41598-018-34820-x

256. Batchelor TT, Gerstner ER, Emblem KE, et al. Improved tumor oxygenation and survival in glioblastoma patients who show increased blood perfusion after cediranib and chemoradiation. *Proc …*. 2013;110(47):19059-19064. doi:10.1073/pnas.1318022110

257. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 Bias Correction. *IEEE*

Trans Med Imaging. 2010;29(6):1310-1320. doi:10.1109/TMI.2010.2046908

258. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30(9):1323-1341. doi:10.1016/j.mri.2012.05.001
259. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 9901 LNCS. ; 2016:424-432. doi:10.1007/978-3-319-46723-8_49
260. Gamer M, Lemon J, Fellows I, Singh P. Various Coefficients of Interrater Reliability and Agreement. *Http://CranR-ProjectOrg/Web/Packages/Irr/IrrPdf*. 2012. doi:https://cran.r-project.org/package=irr%0A
261. Zhou ZH, Wu J, Tang W. Ensembling neural networks: Many could be better than all. *Artif Intell*. 2002;137(1-2):239-263. doi:10.1016/S0004-3702(02)00190-X
262. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Proc 33rd Int Conf Int Conf Mach Learn - Vol 48*. 2016:1050-1059. https://dl.acm.org/citation.cfm?id=3045502. Accessed October 23, 2017.
263. Hoebel K, Chang K, Patel J, Singh P, Kalpathy-Cramer J. Give me (un)certainty -- An exploration of parameters that affect segmentation uncertainty. November 2019. http://arxiv.org/abs/1911.06357. Accessed December 18, 2019.
264. Patel SH, Poisson LM, Brat DJ, et al. T2–FLAIR Mismatch, an Imaging Biomarker for IDH and 1p/19q Status in Lower-grade Gliomas: A TCGA/TCIA Project. *Clin Cancer Res*. July 2017. doi:10.1158/1078-0432.CCR-17-0560
265. Kickingereder P, Sahm F, Radbruch A, et al. IDH mutation status is associated with a distinct hypoxia/angiogenesis transcriptome signature which is non-invasively predictable with rCBV imaging in human glioma. *Sci Rep*. 2015;5:16238. doi:10.1038/srep16238
266. Biller A, Badde S, Nagel A, et al. Improved Brain Tumor Classification by Sodium MR Imaging: Prediction of IDH Mutation Status and Tumor Progression. *Am J Neuroradiol*. 2016;37(1):66-73. doi:10.3174/ajnr.A4493
267. Pope WB, Prins RM, Albert Thomas M, et al. Non-invasive detection of 2-hydroxyglutarate and other metabolites in IDH1 mutant glioma patients using magnetic resonance spectroscopy. *J Neurooncol*. 2012;107(1):197-205. doi:10.1007/s11060-011-0737-8
268. Andronesi OC, Kim GS, Gerstner E, et al. Detection of 2-Hydroxyglutarate in IDH-Mutated Glioma Patients by In Vivo Spectral-Editing and 2D Correlation Magnetic Resonance Spectroscopy. *Sci Transl Med*. 2012;4(116):116ra4-116ra4. doi:10.1126/scitranslmed.3002693
269. Choi C, Ganji SK, DeBerardinis RJ, et al. 2-hydroxyglutarate detection by magnetic resonance spectroscopy in IDH-mutated patients with gliomas. *Nat Med*. 2012;18(4):624-629. doi:10.1038/nm.2682
270. Stadlbauer A, Zimmermann M, Kitzwögerer M, et al. MR Imaging–derived Oxygen Metabolism and Neovascularization Characterization for Grading and *IDH* Gene Mutation Detection of Gliomas. *Radiology*. December 2016:161422. doi:10.1148/radiol.2016161422
271. Zhou H, Chang K, Bai HX, et al. Machine learning reveals multimodal MRI patterns predictive of isocitrate dehydrogenase and 1p/19q status in diffuse low- and high-grade gliomas. *J Neurooncol*. 2019;142(2):299-307. doi:10.1007/s11060-019-03096-0

272. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep*. 2015;5(1):13087. doi:10.1038/srep13087

273. Tang L, Deng L, Bai HX, et al. Reduced expression of DNA repair genes and chemosensitivity in 1p19q codeleted lower-grade gliomas. *J Neurooncol*. June 2018. doi:10.1007/s11060-018-2915-4

274. Akkus Z, Ali I, Sedlář J, et al. Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence. *J Digit Imaging*. June 2017:1-8. doi:10.1007/s10278-017-9984-3

275. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual Deep Convolutional Neural Network Predicts MGMT Methylation Status. *J Digit Imaging*. August 2017. doi:10.1007/s10278-017-0009-z

276. Thomas RK, Baker AC, DeBiasi RM, et al. High-throughput oncogene mutation profiling in human cancer. *Nat Genet*. 2007;39(3):347-351. doi:10.1038/ng1975

277. Cryan JB, Haidar S, Ramkissoon LA, et al. Clinical multiplexed exome sequencing distinguishes adult oligodendroglial neoplasms from astrocytic and mixed lineage gliomas. *Oncotarget*. 2014;5(18):8083-8092. doi:10.18632/oncotarget.2342

278. Wagle N, Berger MF, Davis MJ, et al. High-Throughput Detection of Actionable Genomic Alterations in Clinical Tumor Samples by Targeted, Massively Parallel Sequencing. *Cancer Discov*. 2012;2(1):82-93. doi:10.1158/2159-8290.CD-11-0184

279. Chang K, Zhang B, Guo X, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. *Neuro Oncol*. 2016;18(12):1680-1687. doi:10.1093/neuonc/now086

280. Gorgolewski K, Burns CD, Madison C, et al. Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Front Neuroinform*. 2011;5:13. doi:10.3389/fninf.2011.00013

281. Qi S, Yu L, Li H, et al. Isocitrate dehydrogenase mutation is associated with tumor location and magnetic resonance imaging characteristics in astrocytic neoplasms. *Oncol Lett*. 2014;7(6):1895-1902. doi:10.3892/ol.2014.2013

282. Lasocki A, Tsui A, Gaillard F, Tacey M, Drummond K, Stuckey S. Reliability of noncontrast-enhancing tumor as a biomarker of IDH1 mutation status in glioblastoma. *J Clin Neurosci*. 2017;39:170-175. doi:10.1016/j.jocn.2017.01.007

283. Yamashita K, Hiwatashi A, Togao O, et al. MR imaging-based analysis of glioblastoma multiforme: Estimation of IDH1 mutation status. *Am J Neuroradiol*. 2016;37(1):58-65. doi:10.3174/ajnr.A4491

284. Yu J, Shi Z, Lian Y, et al. Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma. *Eur Radiol*. 2017;27(8):3509-3522. doi:10.1007/s00330-016-4653-3

285. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. ; 2017. doi:10.1109/ICCV.2017.74

286. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*. 2014;15:1929-1958. http://jmlr.org/papers/v15/srivastava14a.html. Accessed April 12, 2017.

287. Akkus Z, Ali I, Sedlář J, et al. Predicting Deletion of Chromosomal Arms 1p/19q in Low-Grade Gliomas from MR Images Using Machine Intelligence. *J Digit Imaging*.

2017;30(4):469-476. doi:10.1007/s10278-017-9984-3

288. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual Deep Convolutional Neural Network Predicts MGMT Methylation Status. *J Digit Imaging*. 2017;30(5):622-628. doi:10.1007/s10278-017-0009-z

289. Chang P, Grinband J, Weinberg BD, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *Am J Neuroradiol*. 2018. doi:10.3174/ajnr.A5667

290. Zhang Z, Chan AKY, Ding X, et al. Glioma groups classified by IDH and TERT promoter mutations remain stable among primary and recurrent gliomas. *Neuro Oncol*. 2017;19(7):1008-1010. doi:10.1093/neuonc/nox042

291. Ene CI, Fine HA. Many Tumors in one: A daunting therapeutic prospect. *Cancer Cell*. 2011;20(6):695-697. doi:10.1016/j.ccr.2011.11.018

292. Mosaic Amplification of Multiple Receptor Tyrosine Kinase Genes in Glioblastoma. 2011:1-8. doi:10.1016/j.ccr.2011.11.005

293. Hu LS, Ning S, Eschbacher JM, et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro Oncol*. 2017;19(1):128-137. doi:10.1093/neuonc/now135

294. Soeda A, Hara A, Kunisada T, Yoshimura SI, Iwama T, Park DM. The evidence of glioblastoma heterogeneity. *Sci Rep*. 2015;5:7979. doi:10.1038/srep07979

295. Freire P, Vilela M, Deus H, et al. Exploratory Analysis of the Copy Number Alterations in Glioblastoma Multiforme. *PLoS One*. 2008;3(12):e4076. doi:10.1371/journal.pone.0004076

296. Hayes J, Yu Y, Jalbert LE, et al. Genomic analysis of the origins and evolution of multicentric diffuse lower-grade gliomas. *Neuro Oncol*. 2018;20(5):632-641. doi:10.1093/neuonc/nox205

297. Lee SY. Temozolomide resistance in glioblastoma multiforme. *Genes Dis*. 2016;3(3):198-210. doi:10.1016/j.gendis.2016.04.007

298. Choi S, Yu Y, Grimmer MR, Wahl M, Chang SM, Costello JF. Temozolomide-associated hypermutation in gliomas. *Neuro Oncol*. February 2018. doi:10.1093/neuonc/noy016

299. Kreisl TN, Zhang W, Odia Y, et al. A phase II trial of single-agent bevacizumab in patients with recurrent anaplastic glioma. *Neuro Oncol*. 2011;13(10):1143-1150. doi:10.1093/neuonc/nor091

300. Friedman HS, Prados MD, Wen PY, et al. Bevacizumab alone and in combination with irinotecan in recurrent glioblastoma. *J Clin Oncol*. 2009;27(28):4733-4740. doi:10.1200/JCO.2008.19.8721

301. Grossmann P, Narayan V, Chang K, et al. Quantitative Imaging Biomarkers for Risk Stratification of Patients with Recurrent Glioblastoma Treated with Bevacizumab. *Neuro Oncol*. 2017:1-32. doi:10.1093/neuonc/nox092

302. Gerstner E, Emblem KE, Chang K, et al. Bevacizumab reduces permeability and concurrent temozolomide delivery in a subset of patients with recurrent glioblastoma. *Clin Cancer Res*. 2019. doi:10.1158/1078-0432.ccr-19-1739

303. Kalpathy-Cramer J, Gerstner ER, Emblem KE, Andronesi OC, Rosen B. Advanced magnetic resonance imaging of the physical processes in human Glioblastoma. *Cancer Res*. 2014;74(17):4622-4637. doi:10.1158/0008-5472.CAN-14-0383

304. Peng J, Zhou H, Tang O, et al. Evaluation of RAPNO criteria in medulloblastoma and other leptomeningeal seeding tumors using MRI and clinical data. *Neuro Oncol*. March

2020. doi:10.1093/neuonc/noaa072

305. Lin NU, Lee EQ, Aoyama H, et al. Response assessment criteria for brain metastases: Proposal from the RANO group. *Lancet Oncol*. 2015;16(6):e270-e278. doi:10.1016/S1470-2045(15)70057-4

306. Donald R, Howells T, Piper I, et al. Forewarning of hypotensive events using a Bayesian artificial neural network in neurocritical care. *J Clin Monit Comput*. 2019;33(1):39-51. doi:10.1007/s10877-018-0139-y

307. Birjandtalab J, Heydarzadeh M, Nourani M. Automated EEG-Based Epileptic Seizure Detection Using Deep Neural Networks. In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE; 2017:552-555. doi:10.1109/ICHI.2017.55

308. Chang PD, Kuoy E, Grinband J, et al. Hybrid 3D/2D Convolutional Neural Network for Hemorrhage Evaluation on Head CT. *AJNR Am J Neuroradiol*. 2018;39(9):1609-1616. doi:10.3174/ajnr.A5742

309. Duong MT, Rudie JD, Wang J, et al. Convolutional Neural Network for Automated FLAIR Lesion Segmentation on Clinical Brain MR Imaging. *AJNR Am J Neuroradiol*. July 2019. doi:10.3174/ajnr.A6138

310. Ay H, Arsava EM, Vangel M, et al. Interexaminer difference in infarct volume measurements on MRI: A source of variance in stroke research. *Stroke*. 2008;39(4):1171-1176. doi:10.1161/STROKEAHA.107.502104

311. Winzeck S, Hakim A, McKinley R, et al. ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. *Front Neurol*. 2018. doi:10.3389/fneur.2018.00679

312. Barber PA, Darby DG, Desmond PM, et al. Prediction of stroke outcome with echoplanar perfusion- and diffusion- weighted MRI. *Neurology*. 1998. doi:10.1212/WNL.51.2.418

313. Wilson JTL, Hareendran A, Grant M, et al. Improving the assessment of outcomes in stroke: Use of a structured interview to assign grades on the modified Rankin Scale. *Stroke*. 2002;33(9):2243-2246. doi:10.1161/01.STR.0000027437.22450.BD

314. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. September 2017. http://arxiv.org/abs/1709.01507. Accessed June 18, 2018.

315. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. June 2016. http://arxiv.org/abs/1606.04797. Accessed September 5, 2018.

316. Pan I, Larson D. Improving Automated Pediatric Bone Age Estimation Using Ensembles of Models from the 2017 RSNA Machine Learning Challenge. *Radiol AI*. 2019. doi:10.1148/ryai.2019190053

317. Saur D, Kucinski T, Grzyska U, et al. Sensitivity and interrater agreement of CT and diffusion-weighted MR imaging in hyperacute stroke. *Am J Neuroradiol*. 2003. doi:10.1901/jaba.1974.7-385

318. Hand PJ, Wardlaw JM, Rivers CS, et al. MR diffusion-weighted imaging and outcome prediction after ischemic stroke. *Neurology*. 2006;66(8):1159-1163. doi:10.1212/01.wnl.0000202524.43850.81

319. Saunders DE, Clifton AG, Brown MM. Measurement of infarct size using MRI predicts prognosis in middle cerebral artery infarction. *Stroke*. 1995;26(12):2272-2276. http://www.ncbi.nlm.nih.gov/pubmed/7491649. Accessed December 14, 2018.

320. Selim M, Fink JN, Kumar S, et al. Predictors of hemorrhagic transformation after intravenous recombinant tissue plasminogen activator: prognostic value of the initial

apparent diffusion coefficient and diffusion-weighted lesion volume. *Stroke*. 2002;33(8):2047-2052. http://www.ncbi.nlm.nih.gov/pubmed/12154261. Accessed December 14, 2018.

321. Thomalla GJ, Kucinski T, Schoder V, et al. Prediction of malignant middle cerebral artery infarction by early perfusion- and diffusion-weighted magnetic resonance imaging. *Stroke*. 2003;34(8):1892-1899. doi:10.1161/01.STR.0000081985.44625.B6

322. Demaerschalk BM, Kleindorfer DO, Adeoye OM, et al. Scientific Rationale for the Inclusion and Exclusion Criteria for Intravenous Alteplase in Acute Ischemic Stroke: A Statement for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke*. 2016;47(2):581-641. doi:10.1161/STR.0000000000000086

323. Mok TCW, Chung ACS. Learning Data Augmentation for Brain Tumor Segmentation with Coarse-to-Fine Generative Adversarial Networks. May 2018. http://arxiv.org/abs/1805.11291. Accessed September 5, 2018.

324. Kohl SAA, Romera-Paredes B, Meyer C, et al. A Probabilistic U-Net for Segmentation of Ambiguous Images. June 2018. http://arxiv.org/abs/1806.05034. Accessed January 21, 2019.

325. Hoebel K, Andrearczyk V, Beers AL, et al. An exploration of uncertainty information for segmentation quality assessment. In: Landman BA, Išgum I, eds. *Medical Imaging 2020: Image Processing*. Vol 11313. SPIE; 2020:55. doi:10.1117/12.2548722

326. Oppenheimer SM, Cechetto DF. Cardiac chronotropic organization of the rat insular cortex. *Brain Res*. 1990. doi:10.1016/0006-8993(90)91796-J

327. Cechetto DF, Wilson JX, Smith KE, Wolski D, Silver MD, Hachinski VC. Autonomic and myocardial changes in middle cerebral artery occlusion: stroke models in the rat. *Brain Res*. 1989. doi:10.1016/0006-8993(89)90625-2

328. Ay H, Koroshetz WJ, Benner T, et al. Neuroanatomic correlates of stroke-related myocardial injury. *Neurology*. 2006. doi:10.1212/01.wnl.0000206077.13705.6d

329. Krause T, Werner K, Fiebach JB, et al. Stroke in right dorsal anterior insular cortex Is related to myocardial injury. *Ann Neurol*. 2017. doi:10.1002/ana.24906

330. Johnson H, Harris G, Williams K. BRAINSFit: Mutual Information Rigid Registrations of Whole-Brain 3D Images, Using the Insight Toolkit. *Insight J*. 2007;(10):1-10. http://hdl.handle.net/1926/1291.

331. Marstal K, Berendsen F, Staring M, Klein S. SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. ; 2016:574-582. doi:10.1109/CVPRW.2016.78

332. Smith SM, Nichols TE. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*. 2009;44(1):83-98. doi:10.1016/j.neuroimage.2008.03.061

333. Woolrich MW, Jbabdi S, Patenaude B, et al. Bayesian analysis of neuroimaging data in FSL. *Neuroimage*. 2009. doi:10.1016/j.neuroimage.2008.10.055

334. Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum Brain Mapp*. 2002. doi:10.1002/hbm.1058

335. Raff E. A Step Toward Quantifying Independently Reproducible Machine Learning Research. September 2019. http://arxiv.org/abs/1909.06674. Accessed January 1, 2020.