Artificial Intelligence for Detection, Characterization, and Classification of Complex Visual Patterns in Medical Imaging; Applications in Pulmonary and Neuro-imaging

Nabil Ettehadi

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Artificial Intelligence for Detection, Characterization, and Classification of Complex Visual Patterns in Medical Imaging; Applications in Pulmonary and Neuro-imaging

Nabil Ettehadi

Medical imaging is widely used in current healthcare and research settings for various purposes such as diagnosis, treatment options, patient monitoring, longitudinal studies, etc. The two most commonly used imaging modalities in the United States are Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Raw images acquired via CT or MRI need to undergo a variety of processing steps prior to being used for the purposes explained above. These processing steps include quality control, noise reduction, anatomical segmentation, tissue classification, etc. However, since medical images often include millions of voxels (smallest 3D units in the image containing information) it is extremely challenging to process them manually by relying on visual inspection and the experience of trained clinicians. In light of this, the field of medical imaging is seeking ways to automate data processing.

With the impressive performance of Artificial Intelligence (AI) in the field of Computer Vision, researchers in the medical imaging community have shown increasing interest in utilizing this powerful tool to automate the task of processing medical imaging data. Despite AI's

significant contributions to the medical imaging field, large cohorts of data still remain without optimized and robust AI-based tools to process images efficiently and accurately.

This thesis focuses on exploiting large cohorts of CT and MRI data to design AI-based methods for processing medical images using weakly-supervised and supervised learning strategies, as well as mathematical (and/or statistical) modeling and signal processing methods. In particular, we address four image processing problems in this thesis. Namely: 1) We propose a weakly-supervised deep learning method to automate binary quality control of diffusion MRI scans into 'poor' and 'good' quality classes; 2) We design a weakly-supervised deep learning framework to learn and detect visual patterns related to a set of different artifact categories considered in this work, in order to identify major artifact types present in dMRI volumes; 3) We develop a supervised deep learning method to classify multiple lung texture patterns with association to Emphysema disease on human lung CT scans; 4) We investigate and characterize the properties of two types of negative BOLD response elicited in human brain fMRI scans during visual stimulation using mathematical modeling and signal processing tools.

Our results demonstrate that through the use of artificial intelligence and signal processing algorithms: 1) dMRI scans can be automatically categorized into two quality groups (*i.e.*, 'poor' vs 'good') with a high classification accuracy, enabling rapid sifting of large cohorts of dMRI scans to be utilized in research or clinical settings; 2) Type of the major artifact present in 'poor' quality dMRI volumes can be identified robustly and automatically with high precision enabling exclusion/correction of corrupt volumes according to the artifact type contaminating them; 3) Multiple lung texture patterns related to Emphysema disease can be automatically and robustly classified across various large cohorts of CT scans enabling investigation of the disease through longitudinal studies on multiple cohorts; 4) Negative BOLD responses of different categories can

be fully characterized on fMRI data collected from visual stimulation of human brain enabling researchers to better understand the human brain functionality through studying cohorts of fMRI scans.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This work was supported by NIH R01-HL121270-05, NIH R01-HL121270-06, NIH R01-HL121270-06S2, NIMH R01-MH110445, NIH UH3OD023328, HHS/NIH R01-MH121070, and NIH R01-MH119510 grants. Data used from the MESA Lung Study is funded by NIH/NHLBI R01-HL130506, R01-HL077612, R01-HL093081, R01-HL112986, RC1HL100543, RD831697, N01-HC-95159, N01-HC-95160, N01-HC-95169, N01-HC-95160, N01-HC-95162, N01-HC-95164, N01-HC-95164, N01-HC-95169, U01-HL114494; and from N01-HC95159 to HC95169. The funding sources for data used from SPIROMICS study is detailed below.

The author thanks the SPIROMICS participants and participating physicians, investigators and staff for making this research possible. More information about the study and how to access SPIROMICS data is available at www.spiromics.org. The author would like to acknowledge the University of North Carolina at Chapel Hill BioSpecimen Processing Facility for sample processing, storage, and sample disbursements (http://bsp.web.unc.edu/). I would like to acknowledge the following current and former investigators of the SPIROMICS sites and reading centers: Neil E Alexis, MD; Wayne H Anderson, PhD; Mehrdad Arjomandi, MD; Igor Barjaktarevic, MD, PhD; R Graham Barr, MD, DrPH; Patricia Basta, PhD; Lori A Bateman, MSc; Surya P Bhatt, MD; Eugene R Bleecker, MD; Richard C Boucher, MD; Russell P Bowler, MD, PhD; Stephanie A Christenson, MD; Alejandro P Comellas, MD; Christopher B Cooper, MD, PhD; David J Couper, PhD; Gerard J Criner, MD; Ronald G Crystal, MD; Jeffrey L Curtis, MD; Claire M Doerschuk, MD; Mark T Dransfield, MD; Brad Drummond, MD; Christine M Freeman,

PhD; Craig Galban, PhD; MeiLan K Han, MD, MS; Nadia N Hansel, MD, MPH; Annette T Hastie, PhD; Eric A Hoffman, PhD; Yvonne Huang, MD; Robert J Kaner, MD; Richard E Kanner, MD; Eric C Kleerup, MD; Jerry A Krishnan, MD, PhD; Lisa M LaVange, PhD; Stephen C Lazarus, MD; Fernando J Martinez, MD, MS; Deborah A Meyers, PhD; Wendy C Moore, MD; John D Newell Jr, MD; Robert Paine, III, MD; Laura Paulin, MD, MHS; Stephen P Peters, MD, PhD; Cheryl Pirozzi, MD; Nirupama Putcha, MD, MHS; Elizabeth C Oelsner, MD, MPH; Wanda K O'Neal, PhD; Victor E Ortega, MD, PhD; Sanjeev Raman, MBBS, MD; Stephen I. Rennard, MD; Donald P Tashkin, MD; J Michael Wells, MD; Robert A Wise, MD; and Prescott G Woodruff, MD, MPH. The project officers from the Lung Division of the National Heart, Lung, and Blood Institute were Lisa Postow, PhD, and Lisa Viviano, BSN; SPIROMICS was supported by contracts from the NIH/NHLBI (HHSN268200900013C, HHSN268200900014C, HHSN268200900015C, HHSN268200900016C,HHSN268200900017C, HHSN268200900018C, HHSN268200900019C, HHSN268200900020C), grants from the NIH/NHLBI (U01 HL137880 and U24 HL141762), and supplemented by contributions made through the Foundation for the NIH and the COPD Foundation from AstraZeneca/MedImmune; Bayer; Bellerophon Therapeutics; Boehringer-Ingelheim Pharmaceuticals, Inc.; Chiesi Farmaceutici S.p.A.; Forest Research Institute, Inc.; GlaxoSmithKline; Grifols Therapeutics, Inc.; Ikaria, Inc.; Novartis Pharmaceuticals Corporation; Nycomed GmbH; ProterixBio; Regeneron Pharmaceuticals, Inc.; Sanofi; Sunovion; Takeda Pharmaceutical Company; and Theravance Biopharma and Mylan.

Dedication

This thesis is dedicated to all immigrants across the globe and the challenges they face every day through the immigration process. From graduate students leaving their home, families, friends, and comfort behind to pursue a better education, to refugees who have been dislocated because of their religion, ideology, sexual orientation, persecution, or war. It is my hope that one day we shall open our eyes to find ourselves living in a world that does not discriminate against people based on their gender, sexual orientation, religion, belief, and the color of their skin. A world of inclusion, kindness, and peace.

Preface

A doctoral program is one of the most challenging yet rewarding academic pursuits. The work in my doctoral journey, documented in this dissertation, would have not been possible without many individuals' collaboration, commitment, guidance, support, and encouragement. Some of whom I will always be indebted to. Thus, I would like to take this opportunity and thank them from the bottom of my heart for making this journey possible.

First and foremost, I would like to thank Dr. Andrew F. Laine, the head of the Heffner Biomedical Imaging Laboratory, for giving me the opportunity to continue my Ph.D. journey under his supervision when I was facing some challenges in my academic career. Due to this opportunity and his mentorship, I was able to work on multiple projects and acquire a wide range of skills. My gratitude extends to you, Dr. Laine. Through your guidance, goodwill, and patience you have taught me many valuable lessons not just for my academic career, but also beyond it. I will always be grateful for that.

My sincere appreciation goes out to Dr. Elsa D. Angelini for her mentorship throughout this process. Her knowledge and experience gave me valuable insight into how to conduct scientific research in the field of applied artificial intelligence, especially in the medical imaging domain. Elsa, your feedback on my research projects and academic writings was always critical and beneficial for my improvement. I would like to thank you for being an important part of this journey.

It is my pleasure to thank Dr. R. Graham Barr, one of our most influential clinical collaborators, for his excellent clinical insight and data accessibility. His critical and constructive

feedback on my work's clinical and technical aspects will always be remembered and appreciated. Thank you, Dr. Barr, for helping me to improve my work and better understand the needs of the clinical community.

I would like to thank Dr. Jonathan Posner, a significant clinical collaborator from Duke University, for his clinical expertise and knowledge. I greatly benefited from the guidance of Dr. Posner on the development and delivery of my work according to the needs of clinicians who use medical images. Thank you, Dr. Posner.

I am grateful for Dr. Qolamreza Ray Razighi's belief in me and his decision to admit me to Columbia University's Department of Biomedical Engineering as a graduate student. In addition, I would like to thank him for his time and effort in mentoring me as my first advisor at Columbia University working in the Quantitative Neuroimaging Laboratory. He taught me a great deal about being very careful and critical when examining engineering methods utilized to process medical data. Ray, thank you for the wonderful journey you took me on at the Quantitative Neuroimaging Laboratory.

My heartfelt gratitude is extended to my talented colleagues and friends at the Heffner Biomedical Imaging Laboratory. It has been a true pleasure working with my dear friend Mr. Soroush Arabshahi. I appreciate his time, his collaboration, and his thoughtful suggestions. Soroush, you have always been generous with your time and patient with me when I had implementation issues related to my work. Thank you so much. I would also like to thank Mr. Xuzhe Zhang for his dedication and close collaboration on multiple projects. Thanks to Mr. Artur Wysoczanski for passing on his experience as the senior student member of the lab, setting up my account, and providing me with access to the data when I joined the lab. These students are

experienced researchers with a very deep understanding of applied artificial intelligence in medical imaging, with whom I had a great deal of fun working together.

Additionally, I would like to take this chance and thank all other collaborators who made this work possible. Many thanks to Dr. Benjamin Smith, Dr. Yun Wang, Dr. Jia Guo, Dr. Pratik Kashyap, and Mr. David Semanek.

I would like to thank the members of my proposal and dissertation defense committees who have not been mentioned so far: Dr. John T. Vaughan Jr for chairing the dissertation defense session, and Dr. Joshua Jacobs for his continuous support. I learned a lot from the course taught by Dr. Jacobs. Lastly, I really enjoyed serving as Dr. Jacob's teaching assistant on the Biostatistics course, working together to minimize the challenges of remote teaching during the early stages of the pandemic.

Lastly, but certainly not least, I want to thank my family for helping me get to this point in my life. As I write this dissertation, it has been nearly seven years since we have been together due to visa difficulties. I would like to thank my mom for teaching me how to be strong to endure the pain of distance and separation. I would like to thank my dad for his continuous support and for teaching me how to tackle life's challenges. Finally, I would like to thank my two sisters for taking care of my mom and dad during my absence.

New York, NY, March 21st, 2022

Nabil Ettehadi

Chapter 1: Introduction

1.1 Medical Imaging

Nowadays, medical imaging serves as a key element of the healthcare system, utilized at all stages such as patient screening, early diagnosis, treatment choice, and follow-up sessions (Bercovich and Javitt 2018). In addition to clinical applications, medical imaging is widely used for research and educational purposes. The currently most popular imaging modalities include Computed Tomography (CT) (Buzug 2011), and Magnetic Resonance Imaging (MRI) (McRobbie et al. 2017) with more than 125 million scans/year (combined) in the US (Papanicolas, Woskie, and Jha 2018). Processing of medical images often includes quality control, artifact detection/reduction, anatomical segmentation of different regions of the body, feature extraction, and recognition and classification of various (potentially damaged) tissue types and complex visual patterns. However, manual processing of such a massive pool of big data (*i.e.*, millions of voxels per image) is significantly labor-intensive and highly subjective. Therefore, it is essential to find optimal automatic approaches for such medical image processing tasks.

1.2 Artificial Intelligence

Artificial intelligence, especially deep learning, has a wide range of applications in today's academic, industrial, healthcare, and commercial domains. This is the result of recent advances in computational resources, as it has become feasible to train large neural networks in a reasonable amount of time for various tasks. The idea behind neural networks is to imitate the structure and our current understanding of the process in which the human brain functions. Specifically, neural networks are composed of numerous individual units (often referred to as neurons), which communicate with each other through weighted connections (similar to interactions between biological neurons in the human brain) (Krenker, Bešter, and Kos 2011; Anaya-Isaza, Mera-

Jiménez, and Zequera-Diaz 2021). The optimal weight values for these connections are learned through the training process to achieve the desired performance for the task at hand (Han et al. 2018). Moreover, neurons in a neural network are activated through a nonlinear activation function, similar to activation in biological neurons of the human brain (Han et al. 2018; Anaya-Isaza, Mera-Jiménez, and Zequera-Diaz 2021).

One area where artificial intelligence and more specifically deep learning becomes very useful is the domain of Computer Vision. With the introduction of Convolutional Neural Networks (CNNs), the field of image processing was revolutionized since CNNs demonstrated outstanding performance (often surpassing human-level performances) in tasks such as image segmentation (Minaee et al. 2021), image classification (Druzhkov and Kustikova 2016), object detection (Zhao et al. 2019), face recognition (Wang and Deng 2021), action recognition (Herath, Harandi, and Porikli 2017), etc. One reason for this is that the introduction of the convolution operator allowed CNNs to be trained on images containing hundreds (or millions) of pixels (or voxels for 3D images) without growing exponentially in parameters (as opposed to a normal neural network). With the proven success of CNNs in the field of computer vision, medical imaging researchers are showing increasing interest in applying them to medical imaging data.

1.3 Artificial Intelligence in Medical Imaging

Recently, Artificial Intelligence (AI) has brought promising applications to the field of healthcare, predominantly in the medical imaging domain (Lakhani et al. 2018). Currently, more than half of all publications on AI in healthcare are related to MRI and CT modalities (Pesapane, Codari, and Sardanelli 2018). Using AI, medical image processing tasks can be efficiently and effectively automated, saving a substantial amount of clinical experts' time and resources. Moreover, AI can identify complex visual patterns that are not necessarily detectable by the human

eye, pushing the medical imaging field away from subjective perceptions to more objective insights (Pesapane, Codari, and Sardanelli 2018). Some examples of AI applications in medical imaging (Liu et al. 2021) include, but are not limited to 1) Classifying medical images into one of several classes of diseases or healthy controls (Yadav and Jadhav 2019; Wang et al. 2020) with applications in dermatology (Esteva et al. 2017; Wu et al. 2020), ophthalmology (Ting et al. 2017; Gulshan et al. 2016), breast cancer (Spanhol et al. 2017), etc.; 2) Detecting and localizing patterns of a certain disease or abnormality (at any stage) from medical images (analogous to the concept of object detection in computer vision) such as detection of lung nodules in chest CT or X-ray images (Liu et al. 2019).; 3) Segmenting medical images into anatomical regions such as lung segmentation on CT (Skourt, El Hassani, and Majda 2018) or brain segmentation on structural MRI images (Zhang et al. 2015; Akkus et al. 2017; Maas, Zabeh, and Arabshahi 2021).

Despite significant contributions of AI to the field of medical image processing (*e.g.*, Ronneberger, Fischer, and Brox 2015; Roy et al. 2019; Ye et al. 2019, in addition to those mentioned above), there remain a set of massive influential imaging datasets for which a reliable AI assistant agent has not been yet developed or lacks optimality, robustness, high precision, and accuracy (Matheny et al. 2019). Our goal in this thesis is to narrow this gap.

1.4. Proposed Artificial Intelligence Approaches in this Thesis for CT and MRI

In this dissertation, we focus on applications of AI in lung CT and brain MR images for detection and classification of complex visual patterns, from artifacts to texture and bio-signals, exploiting scans widely used in popular large cohort studies. Specifically, we will develop AI and mathematical tools to tackle the problems of: 1) Automatic quality control of diffusion MR (Stejskal and Tanner 1965; Le Bihan et al. 1986; Huisman 2003) images, 2) Automatic artifact type classification in diffusion MR images, 3) Automatic lung texture classification for

emphysema disease (Aoshiba, Yokohori, and Nagai 2003; Mets et al. 2012) subtyping, and 4) characterization of under-investigated signals in functional MRI (Ogawa et al. 1992) scans. For solving each problem, a specific strategy from the following three strategies is used:

1) **Weakly-supervised learning:** This strategy utilizes image datasets with weak manual annotations (*i.e.,* binary labels used to denote the presence or absence of an artifact across a 3D whole image) to automate the process of quality control (*i.e.,* problem 1) and different imaging artifact classification (*i.e.,* problem 2) in diffusion MRI volumes. Since manual annotation, especially when dealing with large cohorts of big data, is a highly labor-intensive task, manual annotations at volume-level (instead of voxel/pixel-level) is more practical and convenient for experts in charge of this task. By targeting large cohorts of infant brain diffusion MR images in problem 1 and 2, which will be discussed later in details, we aim to improve child brain development research by providing reliable AI methods for sifting the imaging dataset to make the study results and conclusions more reliable and reproducible.
2) **Supervised learning:** This approach utilizes a set of annotated images (annotated by an algorithm instead of a human expert) for the classification of CT-based spatially-informed lung texture patterns (sLTPs) associated with pulmonary emphysema disease (*i.e.,* problem 3). The supervised annotations, also include locations (3D coordinates) of the present sLTPs (at voxel-level) in the CT image at hand forming a non-weakly labeled dataset. Although the annotations in this approach are at voxel-level, the labels are collected using an algorithm (Yang et al. 2021) instead of expert radiologists. Hence, the labeling load is transferred to machine instead of human. Leveraging the sLTPs, we aim to classify the

novel CT Emphysema Subtypes (CTES) going beyond the three standard emphysema subtypes (Anderson and Augustus et al. 1964; Dahl et al. 2002).

3) **Mathematical modeling and signal processing:** In this thesis, we used appropriate signal processing and statistical techniques in order to detect and characterize different types of Negative Bold Responses (NBRs) in human visual cortex on functional MRI scans (*i.e.,* problem 4).

In this thesis, we use the three strategies mentioned above to tackle the following four specific problems in medical imaging. Each problem is detailed in its corresponding chapter of this thesis which is outlined at the end of this introduction in section 1.5.

1.4.1 Problem 1: Weakly-Supervised Fast Binary Quality Assessment of dMRI

dMRI scans, like any other medical imaging modality, are susceptible to various artifacts. If these artifacts go undetected and be included in the imaging data under investigation, they could damage the result of subsequent analyses. In this problem, we plan to use AI for automatic quality control of dMRI volumes. Specifically, the proposed work aims to develop a weakly-supervised (*i.e.,* binary annotations indicating the quality of a whole 3D dMRI volume, rather than slice or voxel-level delineation) deep learning method for rapid automatic binary quality assessment (*i.e.,* 'poor' vs 'good') of diffusion MRI volumes. This method enables the sifting of a vast amount of raw dMRI scans to avoid damaging the reliability of large cohort studies' results.

1.4.2 Problem 2: Weakly-Supervised Fast Artifact Type Identification in dMRI

In addition to binary quality control of dMRI scans, sometimes it is desired to detect the specific artifact type present in a 3D image. By doing so, the image could be corrected (if possible) for that specific artifact type using the currently available medical image processing tools. This results in reducing the waste in medical imaging data due to complete exclusion of noisy images.

Hence, in order to be ablet to correct the corrupt images (if possible), the first step is to detect the artifact type in the image. Since manual identification of noisy images is highly labor-intensive, in this problem, we aim to leverage AI techniques to automate this process. In particular, the weakly-supervised automated artifact detection framework for diffusion MRI volumes, proposed in this thesis, aims to learn complex visual patterns associated with different artifact types. Using the proposed deep learning framework, the major existing artifacts in whole 3D dMRI volumes can be recognized automatically and rapidly. The corrupt volumes then could be either processed for restoration/quality enhancement according to their type of artifact or completely removed from the dataset to retain the credibility of study results.

1.4.3 Problem 3: Supervised Labeling of Lung Textures Associated with Emphysema Subtypes on CT Images

Recent image processing advances have generated increased interest in automated lung texture analysis on CT images for the *in-vivo* study of lung tissue abnormalities (Depeursinge et al. 2014). Using the unsupervised learning method described in (Yang et al. 2021), ten different emphysema-specific spatially-localized lung texture patterns (sLTPs) were discovered on full-lung high-resolution CT scans in MESA (Bild et al. 2002) and SPIROMICS (Couper et al. 2014) cohorts (details of these two cohorts are presented in Chapter 2). Further evaluation of these sLTPs via statistical data reduction techniques demonstrated that they can be grouped into 6 CT emphysema subtypes that have clinical significance (paper under review). Manual labeling of the discovered 10 sLTPs (associated with 6 CTES) is not only labor-intensive, but also almost not feasible since radiologists are not necessarily able to detect and distinguish these 10 sLTP classes from each other, as some of the classes look very similar to each other by the human eye. Hence, in this problem, we aim to use AI to design a framework for automatic labeling of 10 sLTPs. In particular,

the supervised deep learning framework proposed in this thesis aims to develop a fast, accurate, and robust classifier for labeling 10 sLTPs. It is worth noting that the notion of 'supervised' used here refers to machine-learned ground truth labels provided in (Yang et al. 2021) eliminating the need for manual annotation by radiologists. The proposed deep learning method enables longitudinal studies within and between various large cohorts of lung CT data (*e.g.,* MESA and SPIROMICS) with different scanning protocols for understanding quantitative emphysema and sLTP progression.

1.4.4 Problem 4: Mathematical Characterization of Various Types of Negative Bold Responses in Human Brain on fMRI Scans

fMRI is widely used these days for non-invasive in-vivo imaging of brain functionality. Stimulation of human brain can make local and global changes to the fMRI signal in positive (known as the positive BOLD response) and negative (known as the negative BOLD response) directions (Ogawa et al. 1992). Contrary to the negative BOLD response, the positive BOLD response has been studied widely throughout the literature (Logothetis 2008). Since the relative change in the negative fMRI signal is very small compared to the positive BOLD response and other biological signals in the brain, it is very challenging to detect and characterize this signal. Specifically, one needs to use proper mathematical, statistical, and signal processing tools to ensure the reliability of negative BOLD response characterization. In this problem, we aim to carefully study various types of negative BOLD responses elicited in the human brain via visual stimulation. In particular, the proposed exploratory study in this thesis aims to jointly characterize and investigate the properties of the ipsilateral and contralateral negative BOLD responses due to unilateral visual stimulation observed in human fMRI scans. Linearity analysis, hemodynamic response function extraction, spatial and temporal analysis, and behavioral correlates of the two

negative BOLD responses are investigated to help researchers better understand different types of negative BOLD responses and potentially advance the field toward characterizing their underlying mechanism(s).

1.5 Potential Impact and Thesis Outline

The aim of this work is to leverage and design AI approaches to significantly advance the medical image processing techniques for MRI and CT by: 1) Automatically detecting and classifying visual patterns of artifacts in MR images using weakly-supervised deep learning methods; 2) Automatically and robustly classifying various lung texture patterns (with emphysema correspondence) on human lung CT images across different large cohorts of data; 3) Detecting and characterizing BOLD response patterns from fMRI scans of human brain using proper mathematical, statistical, and signal processing tools.

The proposed weakly-supervised deep learning framework for binary quality assessment of diffusion MR 3D images enables a fast automated method for sifting raw data. This is very important and useful as the inclusion of poor-quality data in any analysis can significantly mislead or bias the result of the study. Using the proposed framework, large cohorts of dMRI data can be rapidly assessed for image quality, eliminating the tedious task of manual visual inspection of the data.

In addition to binary quality control, the proposed weakly-supervised deep learning method for multiclass artifact identification in diffusion MRI volumes allows medical data analysts to automatically detect the type of the major existing artifact in dMRI volumes used in their studies. This is especially useful when discarding a big portion of data due to noisy/corrupt images is not affordable (*e.g.,* in cases where the sample size of a study is relatively small or re-scanning the participants is not an option). In such cases, the images need to be corrected (if possible) according

to their artifact type. Our proposed AI-based framework provides a fast automated approach for sifting large cohorts of data with artifact type labels to be possibly corrected eliminating the need for manual artifact class identification.

The proposed supervised deep learning framework for sLTP classification on lung CT images delivers a fast and robust (*i.e.,* cohort and scanner invariant) labeler for automatic annotation of large cohorts of data with lung texture patterns that are associated with emphysema subtypes. The current understanding of emphysema subtypes and potential ways of slowing the disease progression remains incomplete. Therefore, this automatic labeler may assist researchers in annotating large cohorts of unlabeled (or weakly labeled) lung CT data to be used in studies investigating emphysema.

The proposed exploratory study investigating negative BOLD responses of various types on human fMRI scans provides a clear approach for detecting and characterizing such responses. To the best of our knowledge, to this date, the full properties of various types of negative BOLD responses are not fully explored. In addition, there is no consensus regarding the mechanism(s) responsible for such changes in fMRI signals. Using the results of the exploratory analysis proposed in this thesis, researchers will be one step closer to identifying the possible mechanism(s) underlying the negative BOLD responses in the human brain.

The rest of this thesis is organized as follows: In Chapter 2, we give an overview of the large heterogeneous cohorts of data used in this thesis. Following that, in Chapter 3, we describe the details of the weakly-supervised framework for binary quality assessment of dMRI scans. Next, in Chapter 4, we extend the work implemented in Chapter 3, using a weakly supervised deep learning approach, to identify artifact classes present in corrupt dMRI volumes. Then, in Chapter 5, we will discuss the proposed supervised deep learning method for automatic and robust labeling

of sLTPs related to emphysema disease. Following this, in Chapter 6, we turn our attention to exploring various types of negative BOLD responses elicited in human visual cortex observed in fMRI scans. Finally, in Chapter 7, we present a summary and conclusions of the work done in this thesis.

Chapter 2: Data and Preprocessing

The four problems addressed in this thesis include full lung CT scans, human brain diffusion MRI scans, and human brain functional MRI scans from the following cohorts:

1. SubPopulations and InteRmediate Outcome Measures In COPD Study (SPIROMICS) (Couper et al. 2014) with full lung CT images from visit 1 at full inspiration;
2. Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002) with full lung CT images from visit 5;
3. Adolescent Brain Cognitive Development (ABCD) (Casey et al. 2018) with dMRI human brain images;
4. Healthy Brain Network (HBN) (Alexander et al. 2017) with dMRI human brain images;
5. Columbia University Medical Center (CUMC) in-house fMRI dataset with fMRI scans of human brain.

A summary of the data used from these 5 cohorts is presented in Table 2.1.

Table 2.1: Overview of the medical imaging data used in this thesis.

Study/Dataset	Number of subjects included in our experiments	Number of scans	Number of scanner types	Problem under investigation
SPIROMICS V1 inspiration	2,922	2,922	10	Problem 3 (Chapter 5)
MESA-Lung Exam 5	2,542	2,542	6	Problem 3 (Chapter 5)
ABCD	85	8,530	3	Problems 1 and 2 (Chapters 3 and 4)
HBN	100	12,900	2	Problem 2 (Chapter 4)
CUMC fMRI	42	84	1	Problem 4 (Chapter 6)

For each cohort used in each problem, details of the study, participants (*e.g.*, demographics), and the images (such as the resolution, acquisition parameters, etc.) have been thoroughly discussed in their corresponding chapters. More in-depth information is also available in the corresponding literature of each study as referenced above. It is worth mentioning that for each and every cohort used in this thesis, as described in their respective chapters, their corresponding institutional review boards have approved the design of the study protocols.

The CT images in MESA and SPIROMICS are available at the Heffner Biomedical Imaging Laboratory (HBIL) located in the Department of Biomedical Engineering of Columbia University. The dMRI scans in ABCD and HBN are publicly available online (links are provided in Chapters 3 and 4). At the time of conducting the experiments, the CUMC fMRI study was available at the Quantitative Neuroimaging Laboratory (QNL) of Columbia University Medical Center, but at the time of this writing, it is available at the Quantitative Neuroimaging Laboratory (QNL) of Weill Cornell Department of Medicine[1].

For each problem, a different set of preprocessing steps was carried out. Hence, for more in-depth information on this matter, interested readers are referred to the corresponding chapters to get a detailed overview of all the preprocessing steps conducted for each dataset and problem. In this chapter, we will conclude by providing a brief general overview of the pre-processing steps carried out for each research problem.

Data and pre-processing steps for used in Chapter 3: In this project, we used 85 subjects randomly selected from the pool of subjects in the Adolescent Brain Cognitive Development (ABCD) (Casey et al. 2018) dataset. In particular, we used 8,530 brain dMRI magnitude volumes

[1] https://qnlab.weill.cornell.edu/

(raw images) from all 85 sampled subjects. Each dMRI volume underwent the same exact pre-processing pipeline which includes:

1) Intensity normalization according to the global intensity distribution of all 85 subjects. We used the minimum and maximum intensity values after removing the top and bottom 5% percentiles and linearly normalized the intensity values to fall in the range of [0, 1].
2) Brain extraction using the *dwi2mask* function provided by the MRtrix3 toolbox (Tournier et al. 2019). The brain extraction was carried out to make sure that the deep-learning model focuses on data in brain regions rather than information in the skull which are irrelevant to the goal of the project.

Data and pre-processing steps used in Chapter 4: In this project, we used two datasets. Specifically, we used the exact same portion of the ABCD dataset as well as 100 subjects randomly selected from all of the subjects involved in the Healthy Brain Network (HBN) (Alexander et al. 2017) cohort. The following identical pre-processing steps were applied to both cohorts:

1) Separate intensity value normalizations according to the global intensity information of A) all 85 subjects in ABCD dataset; B) all 100 subjects in the HBN dataset. For each cohort separately, we extracted the minimum and maximum intensity values after eliminating the two 5% percentile tails (i.e., below 5%, and above 95%). Next, we used those values for each cohort separately to linearly normalize their intensity values to be between [0, 1].

Data and pre-processing steps used in Chapter 5: In this project we used two relatively large cohorts of lung CT data from MESA-Lung exam 5 (Bild et al. 2002), and SPIROMICS V1 (Couper et al. 2014) populations. In particular we used CT images from 2,922 distinct subjects in

SPIROMICS and 2,542 distinct subjects in MESA-Lung exam 5. For both cohorts we performed the following pre-processing steps:

1) Lung mask segmentation. For each subject in both cohorts, lung masks were generated using the APOLLO Software provided by VIDA Diagnostic, Iowa. The lung masks were applied to CT images to exclude the non-lung regions.
2) Emphysema and sLTP segmentation. Emphysematous regions within each lung were segmented using the upper limit of normal values. Following the Emphysema segmentation, sLTP ground truth labels were extracted using the method described in (Yang et al. 2021). The sLTP segmented lung CT images were used to form training, validation, and test sets for learning and evaluation process. It is worth noting that there was no training procedure for the MESA-Lung exam 5 cohort as all of the data in this cohort was used only for testing purposes.
3) Intensity normalization. Since intensity values in both cohorts were very similar, we normalized the intensity values in both cohorts using the same minimum and maximum intensity values. Following this procedure, the intensities were linearly normalized to fall within the range of [0, 1].

Data and pre-processing steps used in Chapter 6: In this project, we utilized in-house fMRI scans of 42 subjects. All fMRI scans underwent the following pre-processing steps using FSL (Smith S.M. et al. 2004) and in-house-developed software packages:

1) Slice-timing correction. All fMRI volumes were corrected for the slice acquisition delays using the slice-timing correction methods described in (Parker, Liu, and Razlighi 2017; Parker and Razlighi 2019).

2) Motion correction. For each volume, motion parameters were estimated with respect to the first volume via FSL, using rigid-body transformation. Following the estimation of these parameters, they were applied to the slice-timing corrected volumes to get motion-corrected fMRI volumes.
3) Susceptibility-induced distortion correction. Since for each experiment there were two fMRI scans with opposite phase encoding directions, we were able to estimate the geometric distortion correction field using a susceptibility-induced distortion correction technique called topup (Andersson, Skare, and Ashburner 2003) included in FSL software package.
4) First-level analysis. We modeled the pre-processed fMRI data with 6 predictors for all stimulations and subject responses (i.e., visual right, visual left, audio right, audio left, motor-response left, motor-response right). Predictors were obtained by convolving the boxcar timings with the canonical double-gamma hemodynamic response function. We then used multiple regression to perform the first-level analysis via an in-house-developed software package in Python.

Chapter 3: Weakly-Supervised Learning for Quality Control of Brain Diffusion MRI[1]

3.1 Introduction

Diffusion tensor imaging (DTI) (Basser and Jones 2002; Alexander et al. 2007) is a popular method for studying neural and structural changes in the human brain through Magnetic Resonance Imaging (MRI). In addition to detecting normal changes in white matter (Hüppi and Dubois 2006; Ladouceur et al. 2012; Simmonds et al. 2014; Krogsrud et al. 2016), DTI can be used to detect abnormal conditions such as Alzheimer's disease (Hoy et al. 2017; Lo Buono et al. 2020), and Schizophrenia (Roalf et al. 2015; Tønnesen 2018). Moreover, popular large cohort studies such as the Human Connectome Project (Van Essen et al. 2012) and Adolescent Brain Cognitive Development (ABCD) (Casey et al. 2018) employ DTI data for measuring various biomarkers. DTI data, however, are often subject to MRI artifacts, such as motion, ghosting, Eddy currents, chemical shift, gradient distortions, etc., which could affect the quality and reliability of the results (Ettehadi et al. 2021). Hence, it is imperative that the medical images in studies using DTI are checked for the presence of such artifacts to ensure reliable and reproducible results. Therefore, quality assessment (QA) of DTI volumes is a vital step before any further analysis can be performed (Ettehadi et al. 2021).

Traditional approaches for QA of DTI data have involved the opinion of an MRI expert (or sometimes the consensus opinion of a group of MRI experts) by visually inspecting all volumes (in some cases all individual slices) and looking for any patterns of artifacts (Ettehadi et al. 2021). This task is very challenging and labor-intensive, especially for large cohorts of data, such as the

[1] Disclaimer: Parts of this chapter were taken from (Ettehadi et al. 2021) Copyright © 2021, IEEE.

ABCD dataset. In addition, this approach to QA is subjective, which can lead to disagreements among experts on deciding the quality of an image (Ettehadi et al. 2021). This necessitates automated robust approaches for QA of large DTI datasets.

Fast automated quality control for diffusion MR images is possible with computerized QA of data. A number of works, such as FSL (Jenkinson et al. 2010; Andersson et al. 2016; Bastiani 2019), DTIPrep (Oguz et al. 2014), and DtiStudio (Jiang et al. 2006), have attempted to incorporate some level of automatic quality control in their artifact correction tools. In general, such tools rely on a variety of statistical methods to detect and estimate the underlying noise (Roalf et al. 2016). Alternatively, some approaches utilize Artificial Intelligence (AI) techniques to identify a particular class of artifact patterns (*e.g.*, motions and eddy currents) from the image (Iglesias et al. 2017; Alfaro-Almagro et al. 2018; Graham., Drobnjak, and Zhang 2018). There have been several studies that have evaluated these tools' performance in the QA of DTI data, such as (Liu, Zhu, and Zhong 2015) and (Haddad et al. 2019). These studies have reported that such tools are mostly limited to detecting the specific type of artifact they were designed for. Furthermore, other AI-based approaches that take into account a wider variety of artifacts, such as (Samani et al. 2020), often require slice-wise labels (as opposed to volume-wise labels) in their training set, which is labor-intensive.

In this chapter, we introduce an automatic and fast deep learning tool for identifying 3D DTI volumes of 'good' or 'poor' quality regardless of the type of artifacts present. Contrary to the previous works, our framework relies only on volume-level labels (*i.e.*, a weakly supervised approach) rather than slice/voxel-level labels, which significantly reduces the efforts required for the creation of a manually annotated dataset for training and evaluation. The proposed QA tool consists of two steps. First, we start by using two separately trained 3D Convolutional Neural

Networks (CNNs) in order to classify regions of interest (ROIs) randomly sampled from the entire DTI volume into categories of ‘good’ or ‘poor’ quality. In the second step, the ROI-level labels predicted by the CNNs in the previous step, are then employed in a voting system with two different proposed score functions to predict the final quality label of the whole DTI volume under consideration. The results we report show that our tool is both efficient and practical for performing fast automatic QA of 3D raw DTI volumes, regardless of the class of artifacts present (Ettehadi et al. 2021).

3.2 Method

3.2.1 Overview

In order to assess the quality of a DTI input volume, the proposed framework is structured in two main steps: The first step involves feeding ROIs of different sizes, randomly sampled from the DTI volume under consideration, to two different CNN architectures in order to predict the quality labels of these ROIs. In the second step, through a voting process, the final quality label is assigned to each DTI whole volume based on the predicted ROI-level labels from the previous step. The following section outlines both steps in detail. Fig. 3.1 illustrates the overall framework described in this section.

3.2.2 ROI-Level Label Predictions using CNNs

To capture the different artifacts locally and globally in every part of the 3D diffusion MRI volume, we use two separate CNNs with different 3D ROI sizes. Particularly, we use two sets of ROIs for each and every DTI whole volume: 1) a single large, center-cropped ROI and 2) a number of small overlapping ROIs. This approach allows the overall framework to search for both local and global noise patterns within the image by utilizing multiple-sized ROIs. For each ROI type (*i.e.,* large or small), quality labels are predicted separately using the CNN associated with that

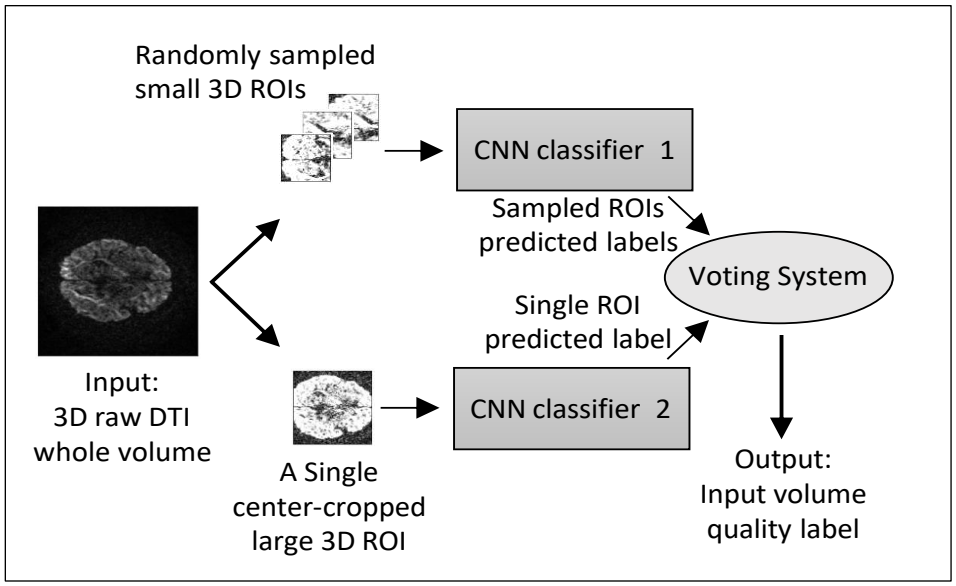

Figure 3.1: Illustration of the proposed two-step automatic whole-volume DTI quality assessment framework. Figure from (Ettehadi et al. 2021) copyright © 2021, IEEE.

ROI type. Fig. 3.2 presents an overview of the architectures of CNN 1 and 2. Each architecture consists of multiple blocks of Convolutional layers, followed by a batch normalization layer. Towards the end of each block, a Max-pooling layer is employed. The image dimension is gradually reduced in both networks while the number of channels is eventually increased. Fig. 3.2 also shows the kernel size and the number of channels for each layer. After the 3rd block, the features are unrolled to form a 128 elements one-dimensional vector. The unrolled features are then fed to a fully connected layer followed by a SoftMax classifier (Goodfellow, Bengio, and Courville 2016) to perform the ROI-level binary quality classification (*i.e.*, class 0: poor quality vs class 1: good quality). In order to regularize the model and minimize potential overfitting effects, a single dropout unit is placed after the fully connected layer.

3.2.3 Whole Volume Label Prediction from ROI-Level Labels via Voting Systems

We utilize the ROI-level predicted labels form step 1 to determine the quality label of the entire 3D DTI volume at hand via a voting block, as shown previously in Fig. 3.1. To this end, we propose two independent tailored score functions for the voting block (each of them can be used

separately) to perform the final classification task. Equation (3.1) defines the score function one which assigns a total score S_{tot} to each volume. S_{tot} is calculated as a linear combination of the predicted label of the single center-cropped large ROI $S_{Center\text{-}cropped\text{-}ROI}$, and the summation of the predicted labels S_i's of N overlapping small ROIs (Ettehadi et al. 2021). The DTI whole volume can be considered good-quality if the total score S_{tot} is above some chosen threshold. Otherwise, the whole volume is classified as a poor-quality one.

$$S_{tot} = \alpha S_{Center-cropped-ROI} + \beta \sum_{i=1}^{N} S_i$$

$$If: \quad S_{tot} \geq threshold \longrightarrow Label = 1 \quad (Good\ volume) \qquad (3.1)$$

$$else: \quad \longrightarrow Label = 0 \quad (Poor\ volume)$$

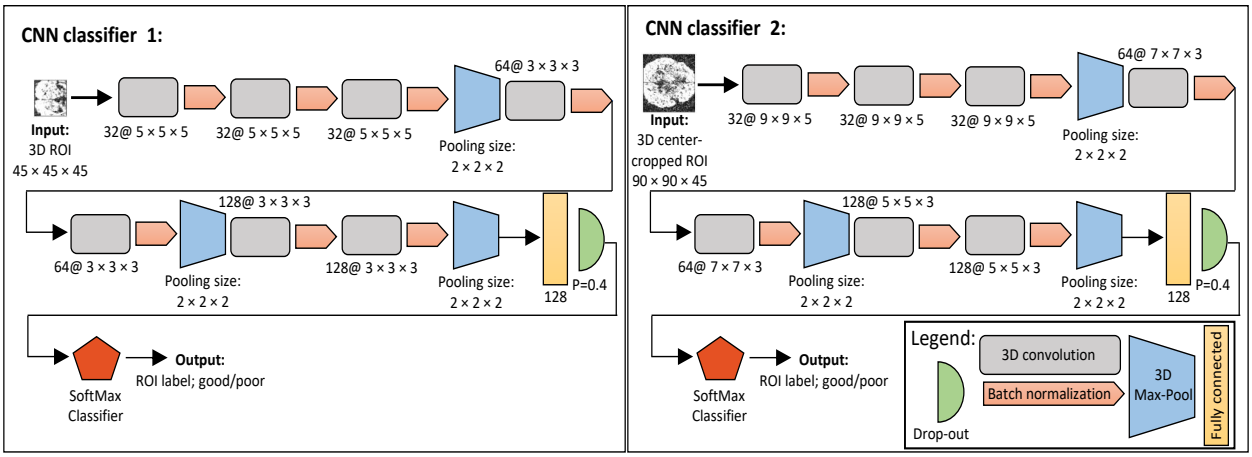

Figure 3.2: The diagrams of the architectures of CNN 1 (left) and CNN 2 (right). After multiple consecutive convolutional blocks followed by Maxpooling, the input ROI is reduced to a feature vector with 128 elements. The SoftMax unit conducts the binary classification taking this feature vector as the input. Figure from (Ettehadi et al. 2021) copyright © 2021, IEEE.

In this voting system, there are 4 hyperparameters, namely α, β, N, and the threshold value. The search space for the choice of these 4 hyperparameters is constrained by the following three conditions: $N \geq 1$, α, β, and threshold ≥ 0, and $\alpha + \beta = 1$. Conversely, voting system 2 is based on a two-tier voting process as in Equation (3.2). In the first layer of this voting procedure, a

thresholding condition is used to assign an aggregate label to the average score of *N* overlapping small ROIs. Next, the DTI whole volume is given a total score S_{tot} by summing the predicted label of the large center-cropped ROI $S_{Center\text{-}cropped\text{-}ROI}$ and the aggregate score of the *N* small ROIs $S_{N\text{-}ROIs}$ calculated earlier. Finally, a decision is made on the quality label of the entire DTI volume in the second layer based on $S_{N\text{-}ROIs}$ and $S_{Center\text{-}cropped\text{-}ROI}$. In particular, the whole DTI volume is considered a good volume if both $S_{N\text{-}ROIs}$ and $S_{Center\text{-}cropped\text{-}ROI}$ are good (*i.e.*, both equal to 1), or else it is assigned a poor-quality label (Ettehadi et al. 2021). This voting system only has two hyperparameters, the threshold value and *N*.

$$S_{AVG-ROIs} = \frac{1}{N}\sum_{i=1}^{N} S_i$$

Layer 1:

$$\begin{aligned} &If: \\ &\quad S_{AVG-ROIs} \geq threshold \;\longrightarrow\; S_{N-ROIs} = 1 \quad (Good\ aggregate\ volume) \\ &else: \\ &\quad \longrightarrow\; S_{N-ROIs} = 0 \quad (Poor\ aggregate\ volume) \\ &S_{tot} = S_{N-ROIs} + S_{Center-cropped-ROI} \end{aligned} \tag{3.2}$$

Layer 2:

$$\begin{aligned} &If: \\ &\quad S_{tot} = 2 \longrightarrow\; Label = 1 \quad (Good\ volume) \\ &else: \\ &\quad \longrightarrow\; Label = 0 \quad (Poor\ volume) \end{aligned}$$

The voting system 1 is further complicated by the presence of its four hyperparameters, which introduce additional freedom to the overall framework to perform more challenging classifications. However, this flexibility comes at the expense of a relatively more rigorous hyperparameter fine-tuning process. Alternatively, as mentioned before, voting system 2 has only

two hyperparameters (*i.e.*, threshold and *N*). This reduces the workload for the fine-tuning process but gives the framework less freedom in determining the final classification results.

3.3 Experiments and Results

3.3.1 Data and Preprocessing

In this work, we used a small portion of the ABCD study (Casey et al. 2018) to train, validate, and test both of our CNN models. Over 10,000 participants were included in the ABCD study between the age of 9 to 10-year-olds and the study tracked brain development from childhood through adolescence. All study procedures were approved by local (corresponding to each site) institutional review boards. Parental informed consent was obtained in writing, and verbal consent was obtained from all children. All of our following experiments were conducted using DTI scans taken from 85 subjects in the ABCD study with multi-shell values (b=0, 500, 1000, 2000, and 3000 s/mm^2) (Ettehadi et al. 2021). All scans have exactly the same resolution (1.7×1.7×1.7 mm^3), the same matrix size (140×140×81), and 96 same diffusion directions. Due to the fact that the scans were acquired across 21 different sites, there is a slight variation in the acquisition parameters reported in Table 3.1.

Table 3.1: Acquisition parameters for all DTI scan used in this work
Table from (Ettehadi et al. 2021) copyright © 2021, IEEE

Number of subjects	Number of b=0 volumes	Number of acquisitions	TE (ms)	TR (ms)
3	8	1	81.9	4100
4	7	1	81.9	4100
7	6	2	89	5300
1	3	1	89	5300
2	8	2	96	6050
1	3	1	96	6050
1	10	2	88	4100
66	7	1	88	4100

An expert with 12 years of experience in MRI analysis manually annotated 8,530 DTI volumes from 85 distinct subjects to facilitate volume-wise classification and significantly reduce labor intensity. Having assigned one single label to each DTI whole volume, the annotated scans form a dataset for weakly-supervised learning. Approximately 5,060 volumes (59.32%) from the labeled cohort were rated as good quality leaving the 3,470 remaining volumes annotated with poor quality labels (Ettehadi et al. 2021). By excluding 1,590 good quality volumes at random, we balanced the two classes in order to avoid running to class imbalance issues when training the two CNNs. In comparison with slice-based QA, volume-based quality assessment is more practical and time-effective since the majority of volumes containing slices of low quality will typically be excluded from any subsequent analysis. Moreover, manual annotations can be done much more easily with volume-wise labeling as it took only around 15 minutes for our expert to label every 100 volumes (corresponding to one subject) (Ettehadi et al. 2021). This is another advantage of the weakly-supervised learning approach.

We shuffle the labeled volumes at random and divide them into train, validation, and test sets with the ratio of 6:2:2. Next, using the global minimum and maximum intensity values, the data intensity of each volume is linearly normalized to the range of [0, 1]. Following this, a brain mask is extracted using the function dwi2mask in the MRtrix3 toolbox (Tournier et al. 2019). During the ROI sampling procedure discussed in the following section, the mask is used to ensure that most of the ROI contains brain areas, rather than background data.

The accuracy, precision, recall, F1-score, and confusion matrix calculated using the predicted labels and the ground truth were used to validate the performance of our tool in assessing the quality of DTI volumes.

3.3.2 ROI Sampling per each DTI Volume

As discussed earlier, each DTI full volume has two sets of sampled ROIs: *N* overlapping small ROIs and a single center-cropped ROI. Specifically in our implementation, we randomly selected 10 overlapping small ROIs of size [45, 45, 45] voxels for each DTI whole volume, as well as a centrally cropped large ROI of size [90, 90, 45] voxels. Overlap between any two small ROIs was limited to a maximum of 52%. Moreover, the sampled small ROIs were allowed a maximum of 16% background voxels in order to capture patterns of artifacts showing up better at the border between background and brain (such as motion).

3.3.3 CNN 1 and 2 Hyperparameters and Training Histories

Fig. 3.2 shows the architectures of CNN 1 and 2. The activation function used for all convolutional layers was RELU (Nair and Hinton 2010), and the kernel parameters were initialized by the random Glorot initialization algorithm (Glorot and Bengio 2010). The convolutional kernel sizes for both CNNs are depicted in Fig. 3.2. For all convolutional layers, the stride size was set to [1, 1, 1] with no padding applied. The stride size of [2, 2, 2] was used for all of the Max-pooling layers. The fully connected layer with 128 elements was activated using the RELU function. In addition, we chose a dropout probability of 0.4. Lastly, the SoftMax layer was activated using the SoftMax activation function, with two possibilities corresponding to the two quality classes indexed as: 0 (poor) and 1 (good).

We implemented both CNN architectures in Python and used Keras (Chollet et al. 2015) with Tensorflow (Abadi et al. 2016) as the backend. Using the manually labeled training dataset discussed earlier in this chapter, the two models were separately trained to minimize the Cross-entropy (Good 1952) loss function. Since the kernel sizes of the two CNNs were different, distinct hyperparameter choices were used to train them individually. Specifically, CNN 1 was trained for

250 epochs with a batch size of 128. We used Adam optimizer with default values (Kingma and Ba 2014) for the first 200 iterations, and SGD with Nesterov optimizer (Sutskever et al. 2013) for the last 50 iterations (initial learning rate = 0.0001, Momentum = 0.5, and decay rate = 10-6). Alternatively, using a batch size of 32, CNN 2 was trained for the entire 250 epochs via SGD with Nesterov optimizer with the same parameter values as in CNN 1. Fig. 3.3 illustrates the training histories of the two CNNs. For both CNNs, the criterion for choosing the best performing model was the highest classification accuracy on the validation set. Fig. 3.4 provides the confusion matrices of the two best-performing CNN models on their respective test sets. CNN 1 achieved ROI-level classification accuracies of 99.85% on the training set, 88.72% on the validation set, and 89.15% on the test set, respectively. As for CNN 2, the accuracies on the training, validation, and test sets were 99.99%, 93.98%, and 94.49%, respectively. In the next step, the framework uses predicted labels from multiple ROIs per each DTI volume to decide the quality label for the DTI volume as a whole, hence there is no need for achieving higher accuracies on ROI-level classification.

3.3.4 Voting Systems for Volume-Level Label Prediction

In section 3.2.3 we presented the logic behind the two proposed voting systems in Equations (3.1) and (3.2). As discussed earlier, voting system 1 includes 4 hyperparameters. In order to have a fair comparison between the two voting systems, we used the same value for N (*i.e.*, number of small ROIs). Specifically, in our implementation, we chose N = 10 which resulted in the best performance on the validation set. Due to the observation that CNN 2 is performing better than CNN 1 (CNN 2 works based on the center-cropped large ROI which captures more regions of the brain making it easier for the model to detect any artifacts), a larger weight is placed

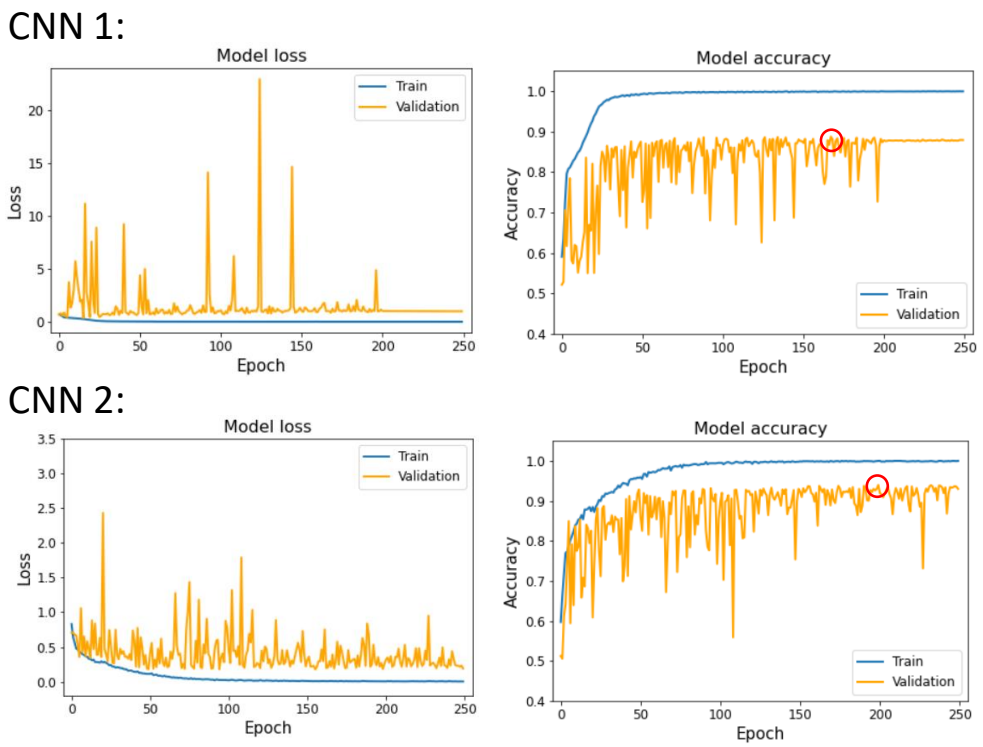

Figure 3.3: Training histories of CNN 1 (top) and CNN 2 (bottom). The histories depict the training and validation loss (left) as well as the training and validation classification accuracies (right) of the two CNNs. The best performing models are marked using red circles. Figure from (Ettehadi et al. 2021) copyright © 2021, IEEE.

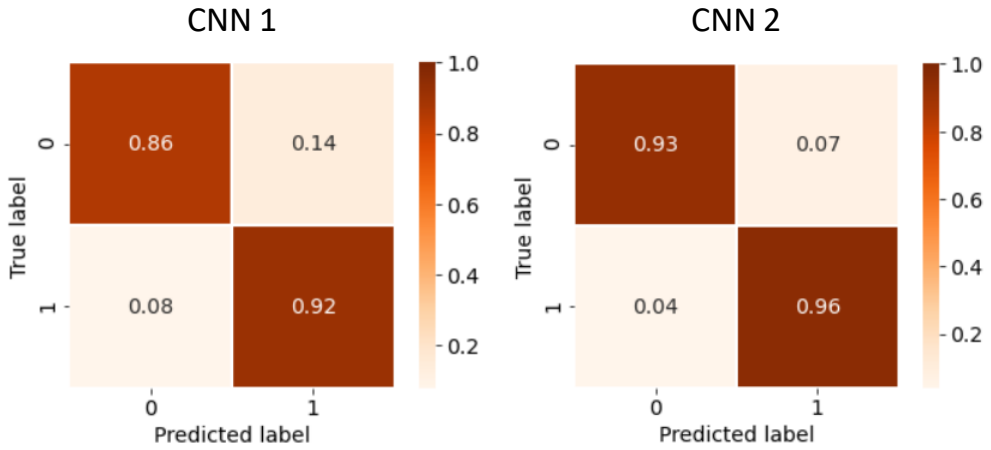

Figure 3.4: ROI-level confusion matrices of CNN 1 (left) and CNN (2) on their respective test sets. CNN 1 results in an average test set classification accuracy of 89.15% while CNN 2 demonstrates a better performance with an average test set accuracy of 94.49%. Figure from (Ettehadi et al. 2021) copyright © 2021, IEEE.

on the scores associated with CNN 2 (*i.e.*, $S_{center\text{-}cropped\text{-}ROI}$). Consequently, we chose $\alpha = 0.6$ and β

= 0.4. The function for calculating S_{tot} presented in Equation (3.1) is a non-continuous function since it takes finite discrete values as inputs. As a result, S_{tot} will also have a finite number of discrete values. Therefore, to find the optimal value for the threshold hyperparameter, we search over the entire possible values of S_{tot} and chose the one corresponding to the best classification performance on the validation set. In Fig. 3.5a, we illustrate this search process by demonstrating the model's total accuracy as well as its true positives and true negatives. As indicated by the black arrow in Fig. 3.5a, the framework achieves the best volume-level classification performance on the validation set over a range of possible threshold values. We chose a threshold value of 2.5 resulting in the same performance between total accuracy, true negatives, and true positives (Ettehadi et al. 2021). In the same way, we varied the threshold value for voting system 2 over the entire space of possible values of the $S_{AVG\text{-}ROIs}$ and selected the one resulting in the highest validation accuracy. Fig. 3.5b illustrates this fine-tuning experiment. The desired model is one that performs as well on both classes (*i.e.*, 'good' and 'poor' quality). As shown in Fig. 3.5b, this voting system achieved high accuracies (although slightly less than that of voting system 1) over a more expansive range of thresholding values compared to voting system 1. This provides evidence that this voting system is more robust to the choice of the thresholding value than voting system 1. The value of 0.1 was chosen resulting in similar behaviors across total accuracy, true positives, and negatives as desired. Next, the chosen threshold values for both voting systems were used to predict the volume-level quality labels on the test set. The result on the test set reported classification accuracy of 100% and 98% for voting frameworks 1 and 2, respectively (99.85±0.15%, and 97.95±0.76% Wilson adjusted accuracy and confidence interval (Wilson 1927)). Moreover, precision, recall, and F1-score were 1, 1, 1 for voting system 1, and, 1, 0.962, 0.981, for voting system 2. Differences in performance between the two voting systems may arise

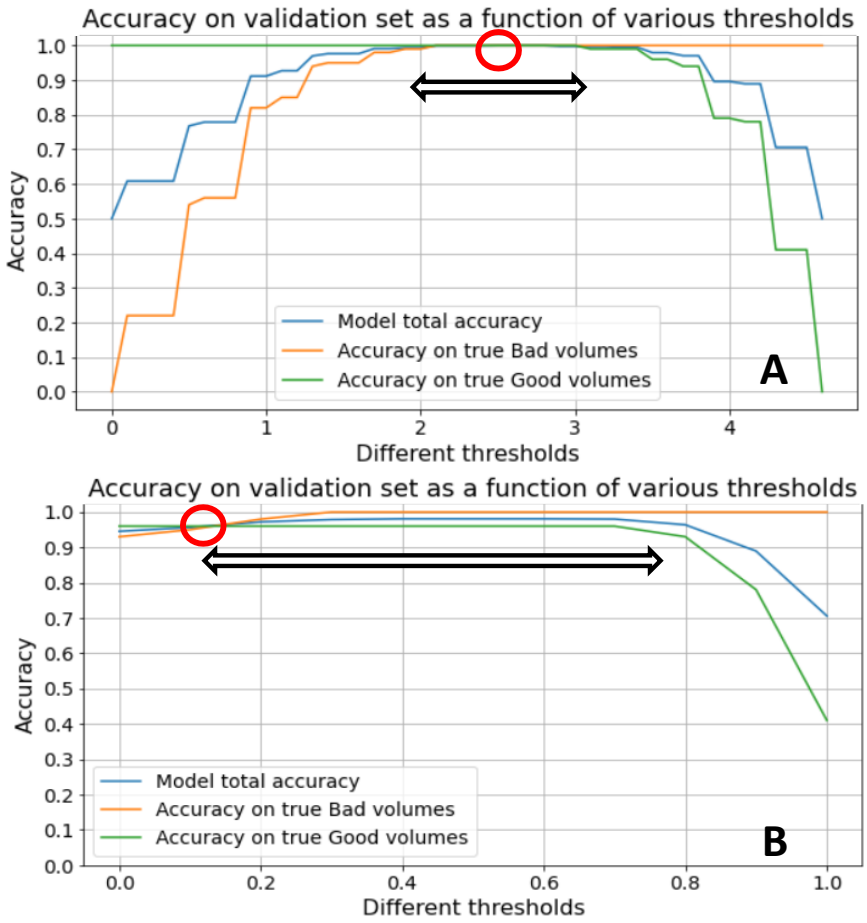

Figure 3.5: Hyperparameters fine-tuning process for (a) voting system 1, and (b) voting system 2 on the validation set. As depicted by the black arrows, voting system 2 shows more robustness to the choice of the thresholding value at the price of being slightly less accurate compared to voting framework 1. Figure from (Ettehadi et al. 2021) copyright © 2021, IEEE.

from situations where individual ROI scores predicted by the two CNNs have a high false-positive rate. In such cases, voting system 1 is more accurate than voting system 2, because, unlike the latter, it does not require that $S_{center\text{-}cropped\text{-}ROI}$ be equal to 1 (*i.e.*, good ROI) to declare the overall DTI volume as good. Utilizing these two voting systems (separately), confusion matrices for volume-level QA of DTI scans in the test set were calculated and shown in Fig. 3.6. Based on these two confusion matrixes, it is evident that our developed deep learning-based quality control tool

is a practical tool for automatic quality assessment of DTI whole volumes, irrespective of the voting system chosen.

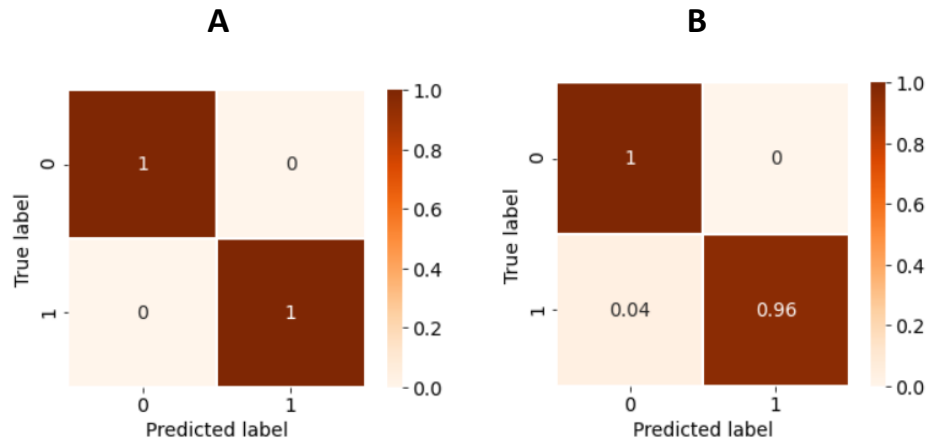

Figure 3.6: Whole volume QA confusion matrices for (a) voting system 1, and (b) voting system 2. Both voting systems result in accurate QA of DTI whole volumes as evident by the insignificant misclassification rates. Figure from (Ettehadi et al. 2021) copyright © 2021, IEEE.

3.4 Discussion and Conclusion

In this chapter, we discussed the details of a developed weakly-supervised learning framework for quality assessment of DTI whole volumes using artificial intelligence. Our framework operates in 2 steps. In the first step, using two carefully designed CNNs of different sizes the quality labels associated with multi-scale ROIs are predicted. The predicted labels are then used in the second step to predict the volume-level quality label through a voting procedure. We proposed two different voting systems performing similarly but with different set of hyperparameters. Our results demonstrated the effectiveness of the proposed framework for automated QA of DTI whole volumes (*i.e.*, 100% and 98% accuracy on the test set using voting systems 1 and 2, respectively) which eliminates the need for manual inspection prior to conducting any subsequent analysis.

Although our model performs well on QA of DTI whole volumes it suffers from artifact class resolution. In other words, the model only provides the user with crude information about the overall quality of a DTI whole volume (*i.e.,* 'good' vs 'bad') and it does not suggest what artifact(s) is present in the volume under investigation or which slices are affected. While in most practices the entire volume identified as a bad quality volume is discarded from subsequent analysis, there are cases where one might want to restore the bad volumes (if possible). Example of such a scenario is when one is conducting research based on a limited number of DTI volumes. In that case, it is not affordable to discard the entire bad volumes that might be restored through some appropriate preprocessing steps. Hence, it is useful to have a model that can tell the user the type(s) of the existing artifact(s) in a bad quality volume (*e.g.,* motion, out of field-of-view, ghosting, etc.). In the next chapter, we propose a solution for this task and detail all of the steps.

Chapter 4: Weakly-Supervised Learning for Multi-Class Artifact Patterns Detection in Brain Diffusion MRI

4.1 Introduction

Diffusion-weighted imaging (DWI) (Stejskal and Tanner 1965; Le Bihan et al. 1986; Huisman 2003; Baliyan et al. 2016) and diffusion tensor imaging (DTI) (Basser and Jones 2002; Alexander et al. 2007) are commonly used techniques today in brain research and clinical neuroimaging (Huisman 2010). Based on the difference in diffusion rates of water across various brain tissues measured using dMRI, one can gain insights into white matter's development (Hüppi and Dubois 2006; Ladouceur et al. 2012; Simmonds et al. 2014). In addition, dMRI can be used to assess white matter abnormalities such as schizophrenia (Roalf et al. 2015; Tønnesen 2018) and Alzheimer's disease (Hoy et al. 2017; Lo Buono et al. 2020). Furthermore, in large population studies, such as the Human Connectome Project (Van Essen et al. 2012), dMRI data is used for measuring different brain biomarkers. There are, however, many sources of artifacts in dMRI data, including but not limited to motion, Eddy currents, low signal-to-noise ratio (SNRs), corrupted field-of-views (FOV), gradient distortions, susceptibility, Gibbs rings, and others (Le Bihan et al. 2006; Krupa and Bekiesińska-Figatowska 2015). If such artifacts are present in dMRI volumes without any exclusion or correction, their presence may lead to unreliable results in any subsequent analysis (Bammer et al. 2003; Van Dijk, RA, Sabuncu, and Buckner 2012; Reuter et al. 2015). It is therefore important to conduct quality control and artifact detection before analyzing dMRI data.

Since manual quality control and artifact identification in dMRI volumes through visual inspection is a highly labor intensive as well as a subjective task, recently computerized methods have emerged to alleviate this challenge. In particular, a number of tools have been proposed to find a level of automation for quality control of dMRI over the years, including FSL (Jenkinson et

al. 2010; Andersson et al. 2016; Bastiani 2019), DTIPrep (Oguz et al. 2014), DtiStudio (Jiang et al. 2006), and TORTOISE (Pierpaoli et al. 2010). More recent advances in dMRI quality control and artifact detection have been made possible by various statistical (Roalf et al. 2016) and artificial intelligence (Kelly et al. 2017; Iglesias et al. 2017; Fantini et al. 2018; Alfaro-Almagro et al. 2018; Graham., Drobnjak, and Zhang 2018; Samani et al. 2020; Ettehadi et al. 2021; Ahmad et al. 2021) approaches. Despite this, most of these approaches either work only on a binary level (*i.e.,* distinguishing between 'poor-quality' and 'good-quality' data without identifying the specific artifact category) (Samani et al. 2020; Ettehadi et al. 2021; Ahmad et al. 2021), or are designed to detect one particular type of artifact (*e.g.,* motion) (Kelly et al. 2017; Iglesias et al. 2017; Fantini et al. 2018). An in-depth analysis of the performance of these tools can be found in (Liu, Zhu, and Zhong 2015; Haddad et al. 2019).

In this chapter, we propose a weakly-supervised deep learning-based framework for automatic multiclass artifact classification in poor-quality dMRI volumes. As opposed to previous works, our method is not tailored to one specific type of artifact and identifies a broader range of artifacts. More precisely, we consider the problem of classifying four categories of artifacts namely: motion, out of field of view (FOV), low signal to noise ratio (SNR), and MRI miscellaneous artifacts. All other MRI artifacts that do not necessarily fall into one of the other three classes fall into the MRI miscellaneous artifacts category, which serves as a control group for the classifier. The proposed method is based on two steps. First, dMRI volumes are divided into mutually exclusive collectively exhaustive (MECE) slabs, which are then fed into a designed convolutional neural network (CNN). As a result, the slabs are labeled according to their artifact class identified by the developed CNN. In the second step, the final artifact label for the dMRI whole volume is then decided via a proposed voting algorithm using the predicted slab-level labels.

Our approach was tested across two datasets separately. Our results indicate that the proposed framework can be used to rapidly classify the four artifact categories under consideration.

4.2 Method

4.2.1 Overview

The proposed framework applies a two-step process to classify dMRI volumes of poor quality into four types of artifacts, as described in the introduction section. The first step involves designing and training a residual squeeze and excitation (SE) CNN to predict the dominant artifact class labels associated with individual MECE 3D slabs extracted from dMRI whole volumes. Next, in the second step, a voting algorithm is proposed based on the rationale of selecting the most consistent artifact type from the predicted slab labels to be reported as the prominent artifact class detected in the dMRI volume at hand. A diagram illustrating the proposed multiclass artifact identification framework is shown in Fig. 4.1. In what follows we discuss the two steps of the proposed framework in detail.

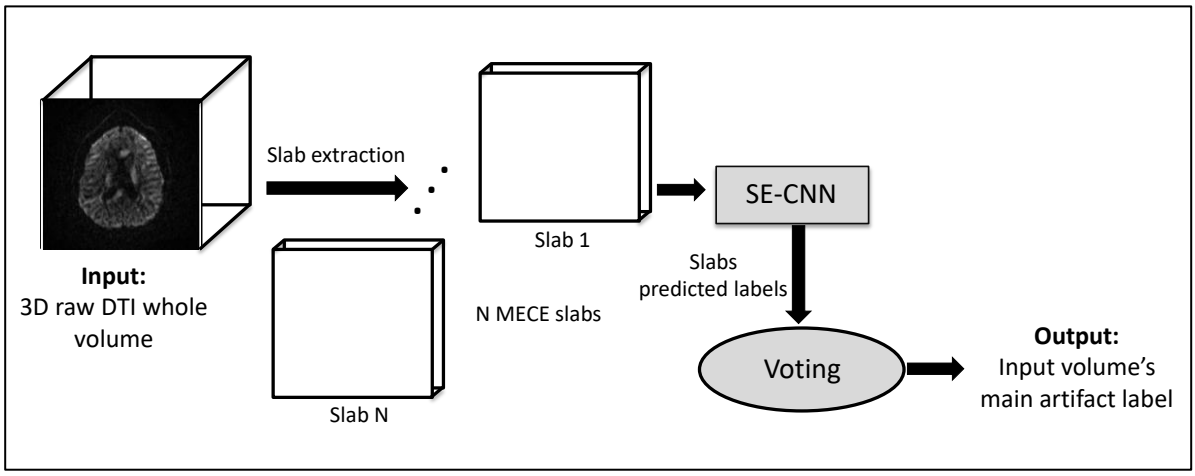

Figure 4.1: An overview of the proposed 2-step multiclass artifact classifier. The 3D input volume is divided into *N* MECE slabs and fed to a residual SE-CNN classifier. The resulting labels are then fed to the voting block to output the final artifact class of the input volume. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

4.2.2 Slab-Level Artifact Label Prediction using a Residual SE-CNN

Given the limited memory capacity of the currently marketed GPUs, it may not always be feasible to feed an entire 3D dMRI volume into a CNN. Thus, partitioning images is necessary before they can be fed to a GPU. To overcome this difficulty while capturing artifact patterns across the entire 3D image without missing any part, we partition the dMRI volumes into MECE slabs covering the entire volume. This technique, in addition to addressing GPU memory limitations, has two additional advantages: a) by scanning the entire volume through 3D MECE slabs, the residual SE-CNN model is able to extract local information while achieving global coherence, and b) it allows for the use of a voting methodology to predict the artifact type in a whole dMRI volume, thereby making the predictions more robust to slab-level errors.

We use a custom-designed residual SE-CNN structure to identify the main type of artifact in each slab within a dMRI volume. Using a number of stacked modified versions of original residual blocks (He et al. 2016) equipped with squeeze and excitation components (Hu, Shen, and Sun 2018), we have designed a custom residual SE-CNN architecture. Fig. 4.2 (left) shows the SE block, while Fig. 4.2 (right) shows the modified residual block upgraded with an SE unit. An SE block represents the interdependencies between the different channels within layers by modeling their interactions (Hu, Shen, and Sun 2018). To put it another way, SE blocks learn how to select the most relevant features for the classification task under investigation by filtering information across channels. Thus, one can think of an SE block as a self-attention mechanism operating on the channel dimension (Hu, Shen, and Sun 2018). The SE block achieves its goal by learning the relative importance weights associated with different channels. To learn the channel weights, a global average pooling along the image spatial dimensions (*i.e.*, x, y, and z) is performed, followed by two fully connected layers with nonlinear activation functions operating on the channel

dimension. One of the key attributes of a SE block is the squeeze hyperparameter *r* (in the first fully connected layer), which reduces the number of channels by a factor of *r* to determine which representation is most critical for the classification task. Using element-wise multiplication, the learned channel weights are then mapped onto the original feature tensor (see Fig. 4.2, left). The interested reader is directed to (Hu, Shen, and Sun 2018) for additional information on this topic. A SE block is then inserted into the double convolutional path of the modified residual block, which allows the channel weights of the two preceding 3D convolutional kernels to be learned (see Fig. 4.2, right). The resulting block is referred to as the modified residual block with SE and it is used consecutively in the architecture of our classifier.

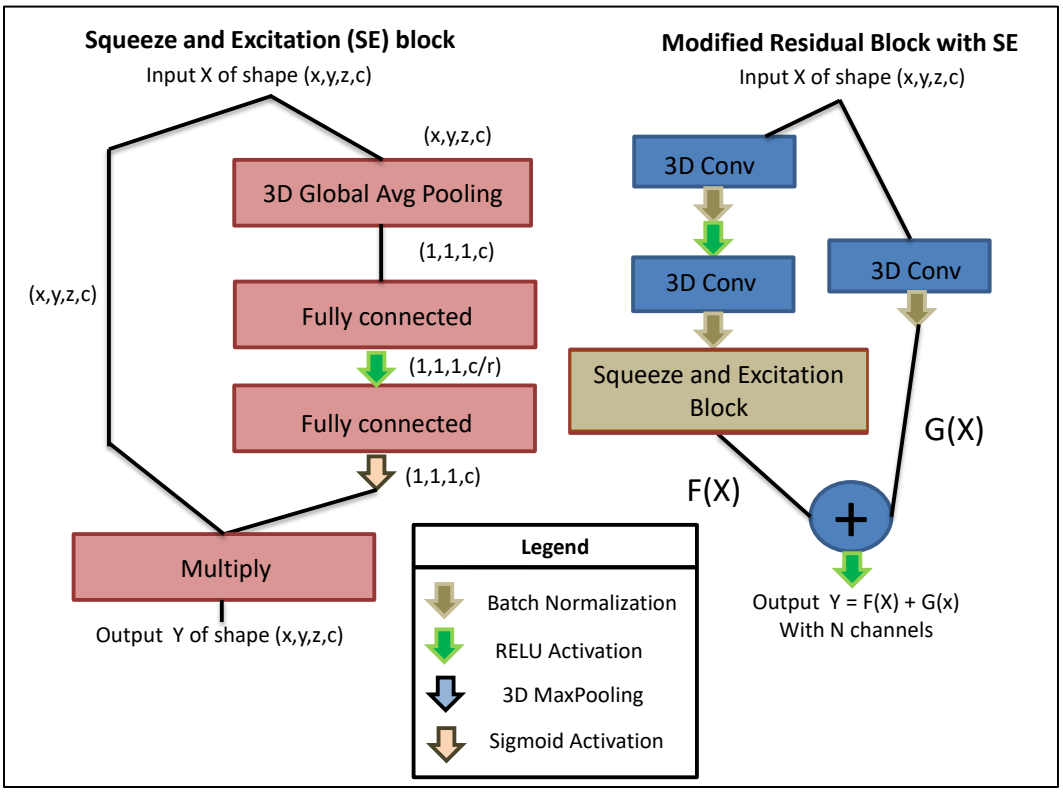

Figure 4.2: Left: The components of an SE block. Right: A modified residual block with SE. A SE block is incorporated into the double convolutional path of the modified residual block to capture relative importance weights for each of the channels of the convolutional filters. This block is the major block used in the architecture of our classifier. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

Fig. 4.3 depicts the architecture of the proposed residual SE-CNN artifact classifier. The residual SE-CNN model is constructed using 5 sequential residual blocks with SE. To reduce the image size after each block, a 3D Maxpooling layer is used. As the network grows deeper, the number of channels is increased while the spatial dimension of the images is reduced. In Fig. 4.3, the kernel sizes and number of channels are shown for each layer. Upon completion of the fifth residual block with SE, the features vector (4,096 elements) is unrolled and fed through two successive fully connected layers. Lastly, the slab-level main artifact type classification is performed via a SoftMax unit with four possible classes (motion, out of FOV, low SNR, and MRI miscellaneous artifacts). As a regularizer and to minimize potential overfitting effects, a dropout unit is added after each fully connected layer. The model has a total of 1,283,029 trainable parameters.

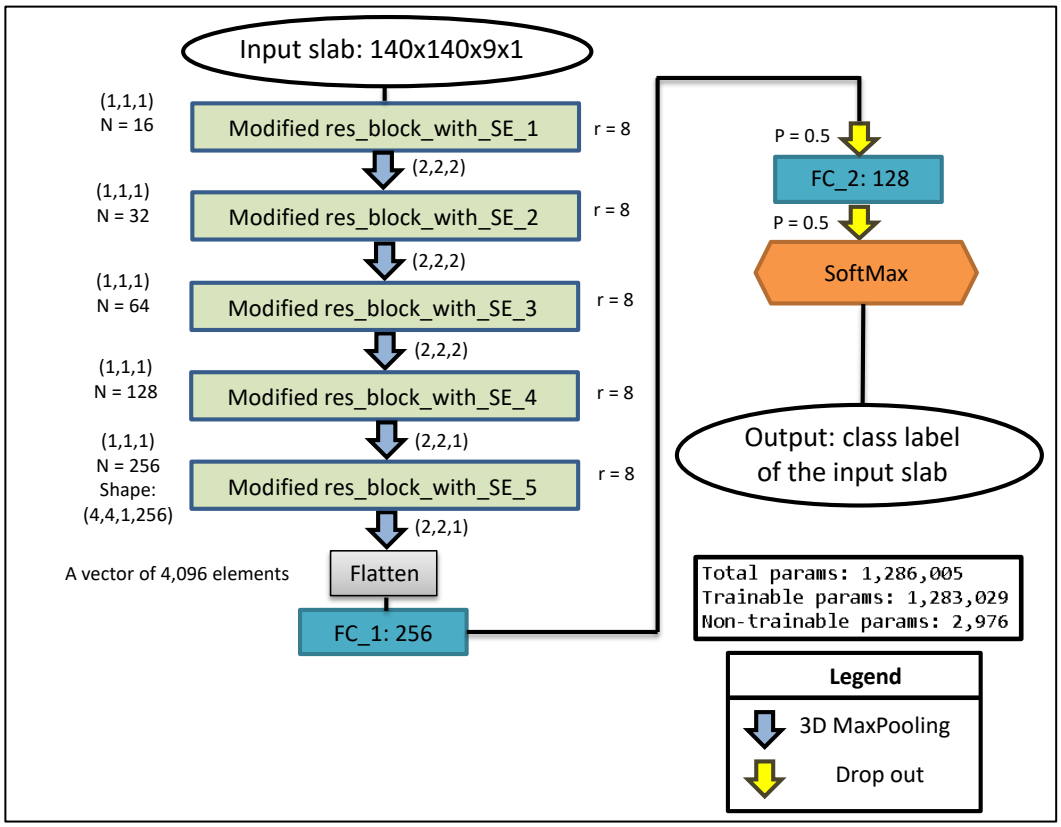

Figure 4.3: An overview of the residual SE-CNN multi-class artifact classifier's architecture. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

4.2.3 Whole Volume Artifact Classification

Following the prediction of the slab-level labels using the residual SE-CNN model, we utilize these labels to identify the dominant artifact label associated with the entire 3D dMRI volume by using the voting block, as previously displayed in Fig. 4.1. The voting block makes a decision on the primary existing artifact type of a dMRI whole volume by taking the predicted labels for the N MECE slabs associated with that volume. Algorithm 1 describes how to achieve this decision through maximum consensus between the predicted labels of the N MECE slabs.

Algorithm 1: Generating whole volume artifact labels

Input: N slabs' predicted labels: $L = \{l_1, l_2, \dots, l_N\}$
Output: Whole volume's predicted artifact label: L_{Vol}

Procedure:

- Let integers 0, 1, 2, and 3 denote the four possible values for each label l_i
- Let n_j denote the count for the j^{th} possible label:
    - Then for each vector L we have:
$$N = \{n_0, n_1, n_2, n_3\}$$
- The whole volume label is predicted as follows:

$$L_{Vol} = argmax\{n_j\} \text{ for j = 0, 1, 2, and 3}$$

4.3 Experiments and Results

4.3.1 Data and Preprocessing

Due to various reasons such as scanner differences, various diffusion directions (multiple gradient directions), diverse demographics, etc., dMRI data tend to be highly heterogeneous. As a result of this heterogeneity, it is nearly impossible to train a CNN on a dataset and use the trained model as is (without changing the learned parameter values) to adjust for differences across studies

(Ahmad et al. 2021). We, therefore, trained our model on two different datasets (separately) and evaluated the results in order to demonstrate the feasibility of automatic artifact classification using the proposed residual SE-CNN architecture. Specifically, we used data from the Adolescent Brain Cognitive Development (ABCD) (Casey et al. 2018) and Healthy Brain Network (HBN) (Alexander et al. 2017) datasets. With separate training for each dataset, the parameters of the proposed residual SE-CNN are optimized according to the target distribution. We describe the two datasets as well as the manual annotation and pre-processing procedures in this section

ABCD dataset: ABCD is a study that is designed to track individual brain development over time (from childhood through adolescence) (Casey et al. 2018). The study recruited more than 10,000 participants aged nine to ten years old to accomplish this goal. The design of this study was approved by the institutional review boards of 21 different sites involved in the study. We used diffusion MR scans from 85 subjects, with b values of 0, 500, 1000, 2000, and 3000 s/mm^2. In all scans, the matrix size is [141×141×81] and the diffusion directions are identical (96). Since the imaging has been performed at 21 different sites, the acquisition parameters are slightly different, which enables the model to learn robustness to heterogeneity in acquisition parameters within the dataset.

HBN dataset: this study was initiated by the Child Mind Institute (CMI, https://childmind.org/) in 2017 (Alexander et al. 2017). The HBN study aimed to create a large dataset of 10,000 children and adolescents aged 5-21 years old living in the New York City area. It recruited participants through community referrals. Chesapeake's Institutional Review Board (https://www.chesapeakeirb.com) approved the study design. Here, we utilized dMRI scans from 100 different subjects with b values of 0, 1000, and 2000 s/mm^2. In all scans the resolution is isotropic (1.8×1.8×1.8 mm^3), there are 72 slices, and the diffusion directions are identical (64). We

resized the HBN images in XY plane using Bi-cubic interpolation since the in-plane matrix size of HBN datasets (mostly 104×104) varies from that of ABCD datasets (140×140). HBN volumes that have been resized end up having a matrix size of [140×140×72].

Since this model is designed to classify the types of artifacts present in poor-quality data, all clean volumes in both ABCD and HBN datasets were excluded from the study. Several tools discussed in the introduction section can be applied to perform binary quality assessment of the data into poor and good quality volumes, such as our AI-based automated quality assessment tool (Ettehadi et a. 2021) discussed in Chapter 3. dMRI volumes with poor quality in both datasets (2,494 in ABCD and 4,226 in HBN) have been manually annotated (weak annotations, *i.e.*, labeling is conducted at the volume level, not the slice level, thus reducing the complexity of manual annotation) into 4 main artifact classes (0 being motion, 1 being out of FOV, 2 being low SNR, and 3 being MRI miscellaneous artifacts), by an expert with 12 years of experience in MRI and DTI analysis. The MRI miscellaneous artifacts class corresponds to any artifacts (such as Eddy currents, ghosting, etc.) that do not fall into any of the other three classes considered here. Among the reasons for having this broad class is the difficulty in identifying a single major source of artifacts in some dMRI volumes. Therefore, this class can be used to train the model to distinguish between the three other prominent artifact types and classify a low-quality volume into a 4th class if it does not fall into the other three. Fig. 4.4 shows the class distributions of manually annotated volumes in ABCD and HBN datasets. A split ratio of 6:2:2 is used to assign the labeled volumes to train, validation, and test sets. Every volume's intensity is normalized to a range of [0, 1]. We did not undertake any brain extraction or background removal in order to capture the visual

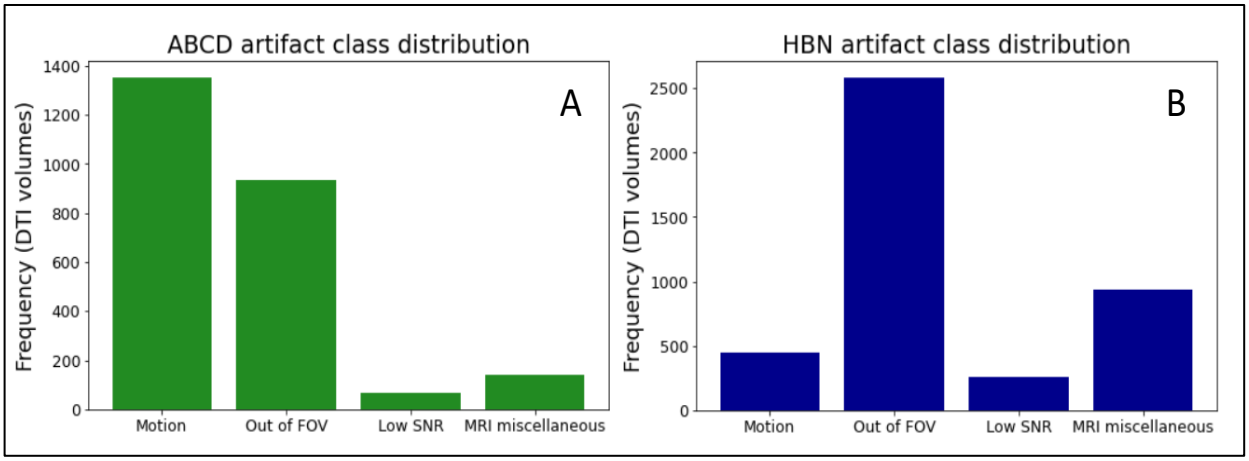

Figure 4.4: Class distribution for the two datasets. A) The ABCD dataset has 1,353 volumes contaminated with motion, 933 volumes being out of FOV, 67 volumes with low SNR, and 141 volumes with MRI miscellaneous artifacts. B) The HBN dataset contains 449 volumes with motion, 2,583 volumes being out of FOV, 258 volumes with low SNR, and 936 volumes with MRI miscellaneous artifacts. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

patterns of the artifacts that exhibit themselves better at the boundary between the brain and background (such as motion and out of FOV).

4.3.2 MECE Slab Extraction from Whole Volumes

All whole dMRI volumes are partitioned into N MECE 3D slabs as discussed in the method section. The original dMRI volume can be formed by connecting these slabs along the z-axis without losing any information. The slab size was set to [140, 140, 9]. Based on this slab size, we are left with 9 slabs per dMRI volume for ABCD and 8 slabs for HBN dataset. These slabs are then fed as inputs to the proposed residual SE-CNN artifact classifier.

4.3.3 Model Hyperparameters

Fig. 4.3 illustrates our proposed residual SE-CNN architecture. In both ABCD and HBN datasets, this particular architecture and its chosen set of hyperparameters were used. The choice of non-linearity was the RELU activation function (Nair and Hinton 2010) for all convolutional filters unless otherwise stated. Glorot initialization (Glorot and Bengio 2010) was used to initialize the kernels' parameters. The size of all convolutional kernels was set to [1, 1, 1] with zero paddings

in order to maintain the same spatial dimension before and after convolutional filtering. The Max-pooling layers used a pooling size of [2, 2, 2] during the first three successive blocks, and [2, 2, 1] during the last two blocks. The reason for this asymmetry in the *z* component of the last two Maxpooling layers is due to the fact that after the 3rd Maxpooling block we are left with a tensor of size [17, 17, 1] with 64 channels. Hence, it is not possible to further reduce the image dimension in the *z* direction. For the SE block's hyperparameter, we set the squeeze parameter r to 8. For the last two fully connected layers (with 256 and 128 nodes, respectively), we applied the RELU activation function. We set both dropout units probabilities to 0.5. Lastly, to perform the slab-level artifact classification from the previous feature vector (*i.e.*, the output of the last fully connected layer with 128 nodes), we used a SoftMax layer (Goodfellow, Bengio, and Courville 2016) with four categories: 0 (motion), 1 (out of FOV), 2 (low SNR), and 3 (miscellaneous MRI artifacts). Using Keras (Chollet et al. 2015) with Tensorflow (Abadi et al. 2016) as the backend, the architecture was implemented in Python. The training hyperparameters for both datasets were, except for epoch number, identical. For the ABCD dataset, the model was trained for 2000 epochs, while HBN provided faster convergence and was trained for 1500 epochs. The batch size of 32 was chosen for both datasets to feed the slabs into the GPU for training. Using the manually labeled data included in each training set, the ABCD and HBN models were trained separately to minimize the Cross-entropy loss function. To optimize the cost function, SGD with Nesterov optimizer (Sutskever et al. 2013) (initial learning rate = 0.0001, Momentum = 0.6, and decay rate = 10-6) was used. The following sections discuss the results of each dataset separately.

4.3.4 ABCD Results

Fig. 4.5 depicts the ABCD training history. Based on the epochs shown here, the training is stable and the overall accuracy is increasing. Model selection was based on the highest overall

classification accuracy on the validation set. Fig. 4.6 indicates the slab-level test set's confusion matrix for the best performing model (at epoch 1999). For the training, validation, and test sets, the model's slab-level classification accuracy was 96.80%, 92.00%, and 91.86%, respectively. We evaluated the performance of the model for the classification of whole volume major artifacts using the voting procedure described in Algorithm 1. On average the model was 96.61% accurate for the classification of the test set's primary artifacts. Fig. 4.7 displays the confusion matrix for whole volume predictions. As depicted, no significant misclassification was observed between the four classes demonstrating the accuracy of the framework in distinguishing between artifacts considered in this work.

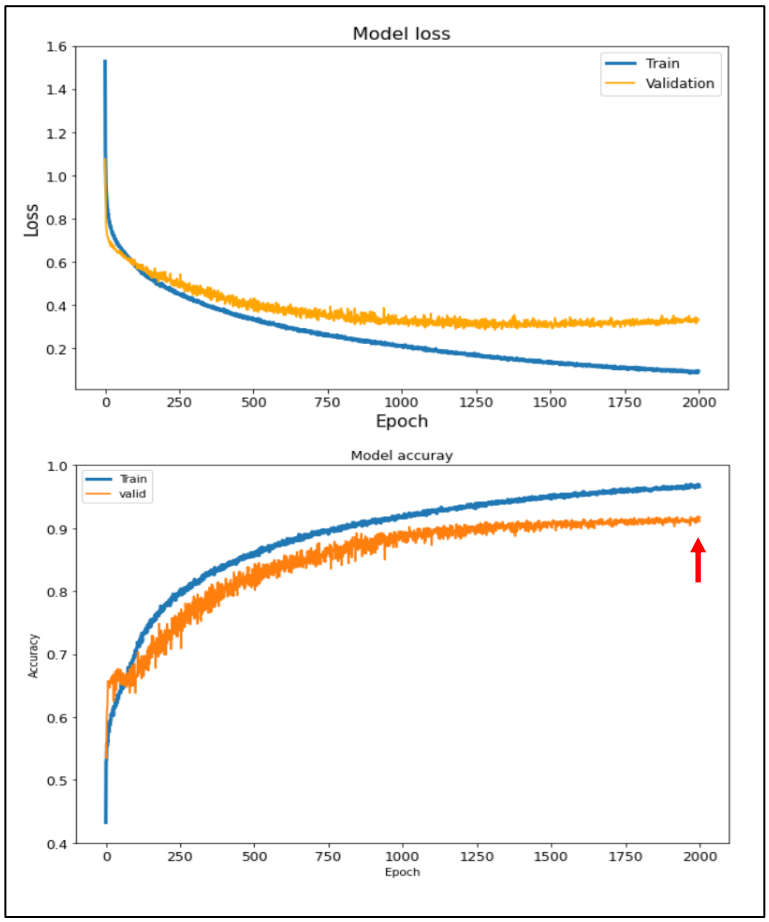

Figure 4.5: Training history on the ABCD dataset. Loss (top), and accuracy (bottom) of the model are shown on the train and validation sets. Best model is achieved at epoch 1999 shown by the red arrow. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

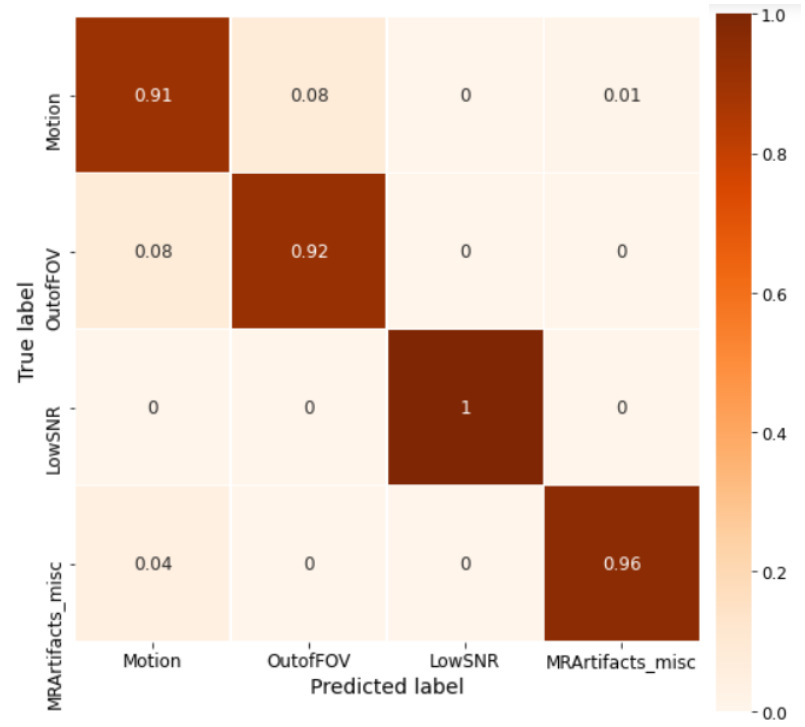

Figure 4.6: Slab-level confusion matrix on the test set of the ABCD dataset. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

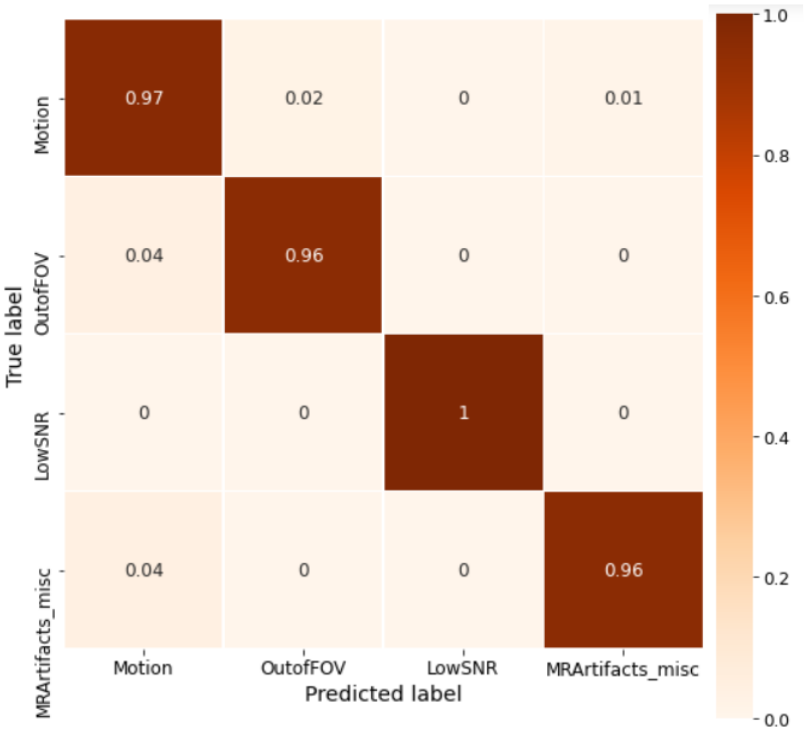

Figure 4.7: Volume-level confusion matrix for ABCD dataset's test set. Using the voting system discussed in Algorithm 1, the model becomes more accurate. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

4.3.5 HBN Results

Using the exact same training hyperparameters as in ABCD, the model was trained for 1500 iterations on the HBN dataset, as shown in Fig. 4.8. The training process for the HBN dataset shows a similar stable behavior as the ABCD dataset. At epoch 1462, the highest validation accuracy model (*i.e.,* the best performing model) led to the slab-level classification accuracy of 95.00%, 94.69%, and 95.77% on the train, validation, and test sets, respectively. The confusion matrix for slab-level artifact classification is shown in Fig. 4.9. It is observed through this confusion matrix, that the classification accuracy is below 90% for the motion class. As the voting process progresses, however, this accuracy will rise, as the individual slab-level misclassification errors will play a less significant role in deciding the volume-level labels. Fig. 4.10 illustrates this concept using the volume-level classification confusion matrix. Overall, the model achieved an accuracy of 97.52% at the volume level, and the accuracy of the motion class significantly improved. Moreover, there is no significant misclassification between the 4 categories of artifacts.

4.4 Discussion and Conclusion

In this chapter, we presented an automatic four-class major artifact classifier for 3D diffusion MRI volumes. The exact architecture was trained with the exact same hyperparameters (both model's and training's hyperparameters) on two different dMRI datasets, namely ABCD and HBN. The trained model provided predictions at slab-level. Using a voting algorithm, slab-level predictions were turned into volume-level artifact class predictions. Based on our results, we demonstrated the ability of the overall proposed framework (*i.e.,* architecture and hyperparameter choices, followed by the voting algorithm) to accurately classify four different categories of artifacts in 3D dMRI volumes.

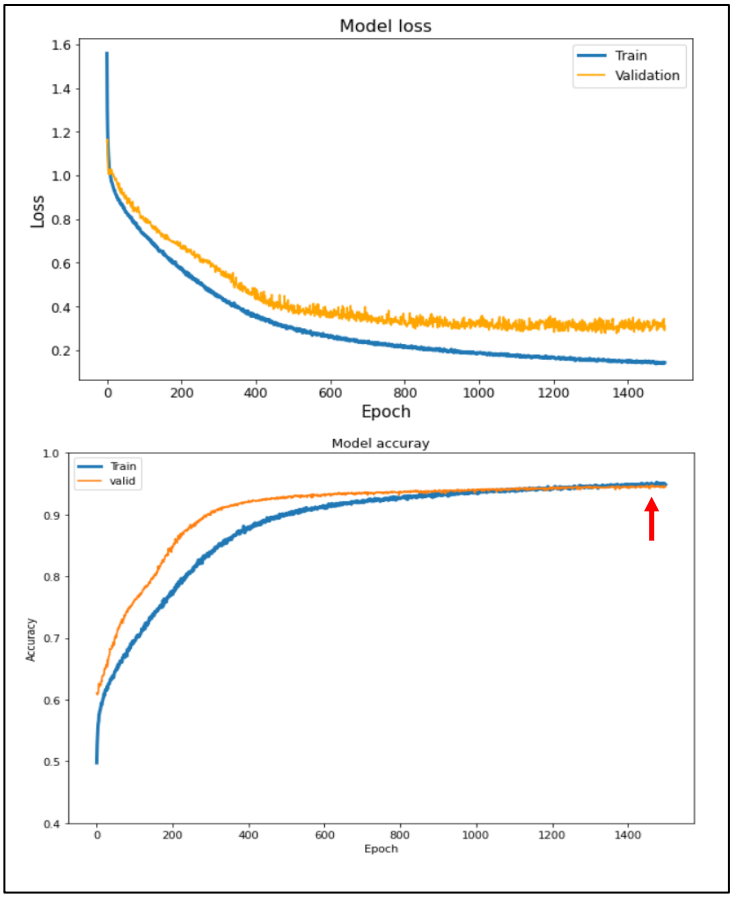

Figure 4.8: Training history on the HBN dataset. Loss (top), and accuracy (bottom) of the residual SE-CNN multiclass artifact classifier are shown on the train and validation sets. The red arrow denotes the best model at epoch 1462. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

As discussed previously, data obtained from diffusion MRI is extremely diverse in nature. For example, an important source of heterogeneity among studies using dMRI data is the difference in diffusion directions. Consequently, designing and training an architecture that generalizes with an acceptable degree of accuracy to unseen and different distributions is extremely challenging. One key reason is that, when trained on a dataset with a specific set of diffusion directions, 3D convolutional kernels learn their spatial parameters' optimal values according to those directions. Thus, when the model is applied to a new dataset with different diffusion directions, the same learned values for the convolutional kernels are not optimal according to the new diffusion directions. Furthermore, in our architecture, the channel attention

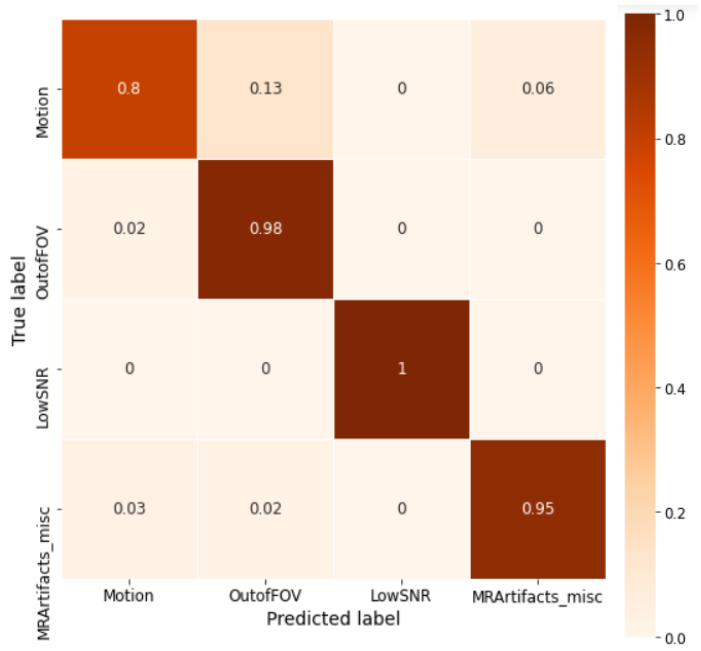

Figure 4.9: Slab-level confusion matrix on the test set of the HBN dataset. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

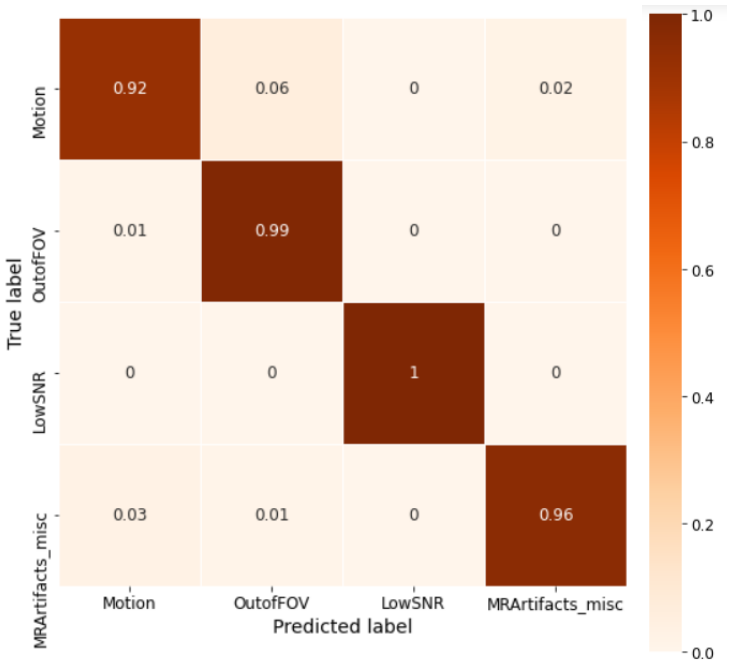

Figure 4.10: Volume-level confusion matrix for HBN's test set. Using the voting system discussed in Algorithm 1, the accuracy on the motion class improved significantly. This demonstrates the importance of the voting process in the final classification. Figure from (Ettehadi et al. 2022) copyright © 2022, Frontiers.

weights learned in SE blocks are also computed according to the diffusion directions of the original

dataset on which they were trained. For a dataset with different diffusion directions, these learned attention weights are not necessarily optimal. As a result, applying a model that is trained on a specific dataset, while fixing its parameter values, to a different dataset is not a viable option.

To address this issue, we proposed a system that includes a set of model's hyperparameters (not learnable parameters, this distinction is important) as well as the hyperparameters for the training process. We showed that the optimal model parameters values for a new dataset (with various sources of heterogeneity, such as different diffusion directions) can be found by training the architecture using a small subset of the data. Studies that examine binary (*i.e.*, 'poor' vs. 'good') quality control of dMRI volumes such as (Ahmad et al. 2021), which is a simpler task than an artifact classification with four classes, have also proposed re-training the network on a small subset of manually annotated data to overcome the generalization problem when faced with a new dMRI dataset.

A neural network's training process often faces two major challenges: 1) the tedious task of manual annotation, and 2) determining the model's and the training's hyperparameters. To ensure potential users can train the model according to their target diffusion MRI datasets, we have attempted to address these two challenges within our proposed framework. First, because our framework is based on a 3D architecture, the manual annotation task can take place only at the volume level (not slice/voxel level) by assigning a single label to an entire volume. In other words, we are using a weakly-supervised method to train the network. As a result, the manual annotation process will be significantly reduced. In our case, for instance, our expert annotated one subject (~110 volumes) with four classes of artifacts in only ~17 minutes. Second, we identified a set of hyperparameters for both the model's architecture and its training process that provided comparable results across two distinct datasets. Considering these findings, we believe that

potential users may be able to use these sets of hyperparameters for training the residual SE-CNN classifier on a different dataset, or at the very least as a base for reducing the domain of search.

As previously discussed, most work on the detection of artifacts or quality control of dMRI scans has been conducted at the binary (‘poor’ vs. ‘good’) level. At the time of this writing, the proposed approach is the first automated and accurate method that considers a broader range of artifact classifications (in particular, 4 classes of motion, out of FOV, low SNR, and MRI miscellaneous artifacts). Obtaining dMRI scans is a labor-intensive and expensive process. It involves a time-consuming scheduling process (especially in research settings) and a great deal of human-to-human interactions. As a result, we believe that by identifying the type of artifact in low-quality volumes using our method, one may be able to correct for the detected artifact (if possible) and restore the volumes in order to lower the cost of acquiring dMRI scans.

Although we demonstrated that our method can be used to automatically classify four categories of artifacts in dMRI volumes of poor quality, we are aware that it is not perfect. Currently, our method is limited to detecting only four types of artifacts. In future work, the framework could be further improved upon to be able to consider a broader range of artifacts. To put it another way, we plan to extensively examine the different types of artifacts associated with MRI miscellaneous (the fourth class). Additionally, each volume is currently identified with only one major artifact type. It might be possible to adapt the model in the future to output multiple artifact types in cases where there are various artifact categories in the volume being examined.

Chapter 5: Supervised Learning and Robust Labeling of Spatially-Informed Lung Texture Patterns Associated with CT Emphysema Subtypes

5.1 Introduction

Emphysema refers to the irreversible loss of lung tissue that does not result from fibrosis (Mets et al. 2012). Particularly, emphysema is defined as the expansion of the airspaces accompanied by the destruction of alveolar walls (Aoshiba, Yokohori, and Nagai 2003). The condition of pulmonary emphysema is a major contributing factor to Chronic Obstructive Pulmonary Disease (COPD), which is considered to be the third leading cause of death worldwide, according to the World Health Organization (World Health Organization 2020). Emphysema and its various types are not fully understood. Traditionally, before the availability of computed tomography (CT) technology, three distinct types of emphysema were identified using a small number of autopsy sessions. Namely, centrilobular (CLE), panlobular (PLE), and paraseptal (PSE) emphysema (Anderson and Augustus et al. 1964; Dahl et al. 2002). Due to their associations with different risk factors and clinical manifestations (Dahl et al. 2002), these subtypes may suggest three distinct lung diseases. Pathologists are, however, not certain whether these three pure subtypes exist (Anderson and Augustus et al. 1964) because of various factors, such as a lack of CT-based lung texture analysis and the small set of autopsy series leading to their discovery. Latest advances in CT technology and computerized lung texture analysis have identified ten groups of Spatially-informed Lung Texture Patterns (sLTPs) on full lung and cardiac CT scans, that have been associated with the traditional three Emphysema subtypes as well as potentially novel Emphysema subtypes (Häme et al. 2014; Häme et al. 2015; Yang et al. 2016a; Yang et al. 2016b;

Yang et al. 2017; Yang et al. 2019; Yang et al. 2021). With the help of unsupervised machine learning techniques, (Yang et al. 2021) have discovered ten sLTPs from MESA-Lung (Bild et al. 2002) and SPIROMICS (Couper et al. 2014) data sets. Additionally, visual inspection of these 10 sLTPs by expert radiologists as well as the use of statistical data reduction techniques have concluded that they cluster into six CT Emphysema Subtypes (CTES) that extend beyond the commonly recognized subtypes (paper under review).

Due to the difficulties associated with labeling lung CT scans manually with multiple texture patterns (*e.g.,* 10 sLTPs) to determine emphysema subtypes, automated computerized labeling methods have recently become popular. A variety of image processing techniques and artificial intelligence approaches, such as supervised machine learning, have been employed to label the traditional emphysema subtypes on lung CT scans, as shown in (Sørensen, Shaker, and Bruijne 2010; Ginsburg et al. 2012; Gangeh et al. 2010; Asherov, Diamant, and Greenspan 2014; Häme et al. 2015; Binder et al. 2016; Yang et al. 2016b; Ross et al. 2017; Bermejo-Peláez et al. 2018; Peng et al. 2018; Peng et al. 2019a; Peng et al. 2019b; Li and Mukundan 2021; Wu et al. 2021). These approaches, however, only addressed the three traditional emphysema subtypes and did not cover the novel 6 CTES or 10 sLTPs. In more recent work, (Yang et al. 2021) classified the 10 sLTPs and 6 novel CTES (paper under review) in the MESA and SPIROMICS cohort lung CT images using machine learning approaches. Despite the high and acceptable accuracy of the sLTP classification reported in (Yang et al. 2021), the approach is not fully robust to variations in cohorts and scanner types as it requires exhaustive parameter tuning or domain adaptation. Hence there is a need for Emphysema classification approaches that are more robust to change in cohorts and scanner types.

In this chapter, we present a supervised deep learning framework for robust classification of the 10 sLTPs leading to 6 CTES. In order to classify the 10 SLTPs, we have developed an attention-based Convolutional Neural Network (CNN). Following the automated labeling of the 10 sLTPs by the proposed CNN, we then utilize the definition of 6 CTES (derived from clustering of 10 sLTPs mentioned in (paper under review)) to classify the novel emphysema subtypes. Using the SPIROMICS inspiratory dataset (Couper et al. 2014), we trained, validated, and tested our model. Our model is also tested on the MESA-Lung exam 5 dataset (Bild et al. 2002) to test its robustness to different cohorts. Additionally, we present the classification accuracy for each scanner type in each cohort. Our results demonstrate that our method can be used to accurately classify sLTPs clustered into CTES in human lung CT images obtained by different scanners from multiple cohorts.

5.2 Method

5.2.1 Overview

The proposed framework employs randomly sampled Regions of Interest (ROIs) automatically labeled by the method described in (Yang et. al 2021). The model takes ROIs of fixed size as inputs, and outputs their sLTP class labels. More specifically, in order to train, validate, and test the proposed attention-based convolutional neural network, ROIs and their labels are used in a supervised learning approach. The following section provides a general overview of our framework for supervised classifications of ROIs based on 10 sLTPs and 6 CTES.

5.2.2 ROI Sampling

We randomly sampled ROIs of size 36×36×36 voxels from all over the lung. This choice of ROI size in SPIROMICS and MESA-Lung cohorts, which results in cubes of approximately $25mm^3$ (similar to the size of secondary pulmonary lobules), was exactly the same as in (Yang et

al. 2017) where the sLTPs were originally discovered. Fig. 5.1 depicts an example of an ROI of this size. An sLTP label (generated by the method in (Yang et al. 2021)) is stored for each ROI to be used for supervised training of the proposed model. More detail on sampling criteria for ROIs is presented in section 5.3.2.

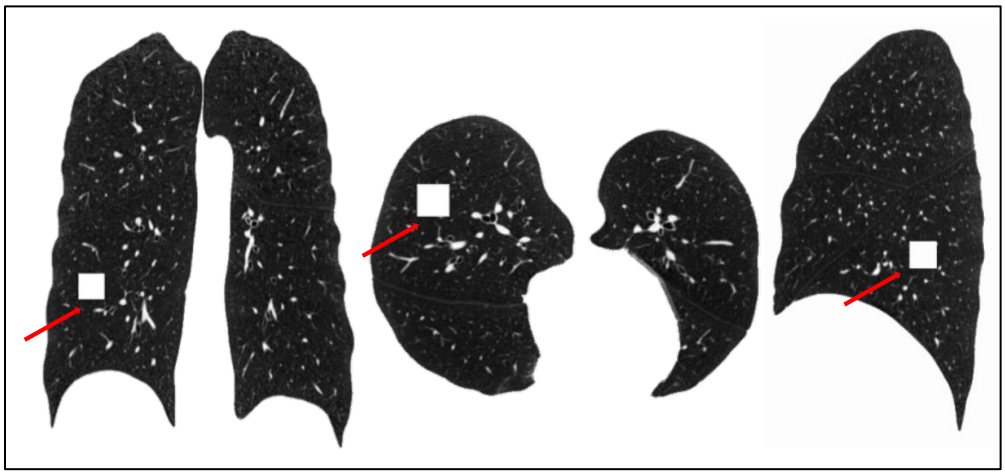

Figure 5.1: Three 2D views of the lung of a subject showing an arbitrary 3D ROI of 36×36×36 voxels. The ROI is highlighted in white and is pointed to by the red arrows.

5.2.3 Residual Squeeze and Excitation Convolution Neural Network sLTP Classifier

In Fig. 5.2, we illustrate the overall architecture of our model, which takes inputs of size 36×36×36 voxels and predicts their sLTP labels. The model is primarily composed of two blocks: 1) residual blocks (He et al. 2016), and 2) squeeze and excitation (SE) blocks (Hu, Shen, and Sun 2018). As depicted in Fig. 5.2a, a modified version of the residual block is utilized in this framework. According to Fig. 5.2a, the residual block contains two branches. On the left, there are two successive 3D convolutional layers with N channels. The first convolutional layer is followed by batch normalization and a RELU activation function (Nair and Hinton 2010). The second convolutional layer involves only a batch normalization layer. The right branch, on the other hand, uses a single 3D convolutional layer with N channels prior to a batch normalization layer. As a final step, the outputs from the two branches are combined and then activated by the RELU

nonlinearity activation function. Fig. 5.2b illustrates the SE block architecture. As mentioned in the previous chapter, SE blocks provide a form of self-attention mechanism designed by (Hu, Shen, and Sun 2018) specifically for CNNs. Essentially, a SE block seeks to find a representation between channels that captures their interdependencies to facilitate the classification task (Hu, Shen, and Sun 2018). In Fig. 5.2b, we see that the SE block is divided into two branches: a) a skip connection branch (left) and b) a branch that determines the optimal channel representation. The latter takes the input feature tensor and passes it through a 3D global average pooling and two fully connected layers. Through the pooling layer, the 3D image is reduced to 1×1×1 while the channel dimension remains intact (the squeeze operation). The first level of fully connected layers reduces the channel numbers by a factor of r in an attempt to find the most important channels for classifying the data. Afterward, a RELU non-linearity is applied. The 2nd fully connected layer restores the channel's number to its original value. During this step, various channels are given different weights, which reflect their importance in the overall classification (the excitation operation). This is followed by a Sigmoid activation function (Han and Moraga 1995). Lastly, the outputs of both branches are multiplied together (element-by-element) to produce a 4D tensor of the same size as the original input to the SE block combined with channel attention weights. As shown in Fig. 5.2c, residual and SE blocks are mixed to form a single block, called the modified residual block with SE. The final deep learning architecture comprises four stacked residual blocks with SE as shown in Fig. 5.2d. For the purpose of reducing the spatial dimensions of the images after each block, we utilize the 3D Maxpooling operator. The features are unrolled following the completion of the last 3D Maxpooling layer, and are then fed to two consecutive fully-connected layers (each followed by a RELU activation function) with 512, and 128 nodes, respectively. We used two Dropout units in order to minimize the effects of overfitting on the fully connected layers.

The sLTP classification is completed by a SoftMax layer with 10 classes. There are a total of 2,914,890 parameters in this model, of which 5,760 are not trainable.

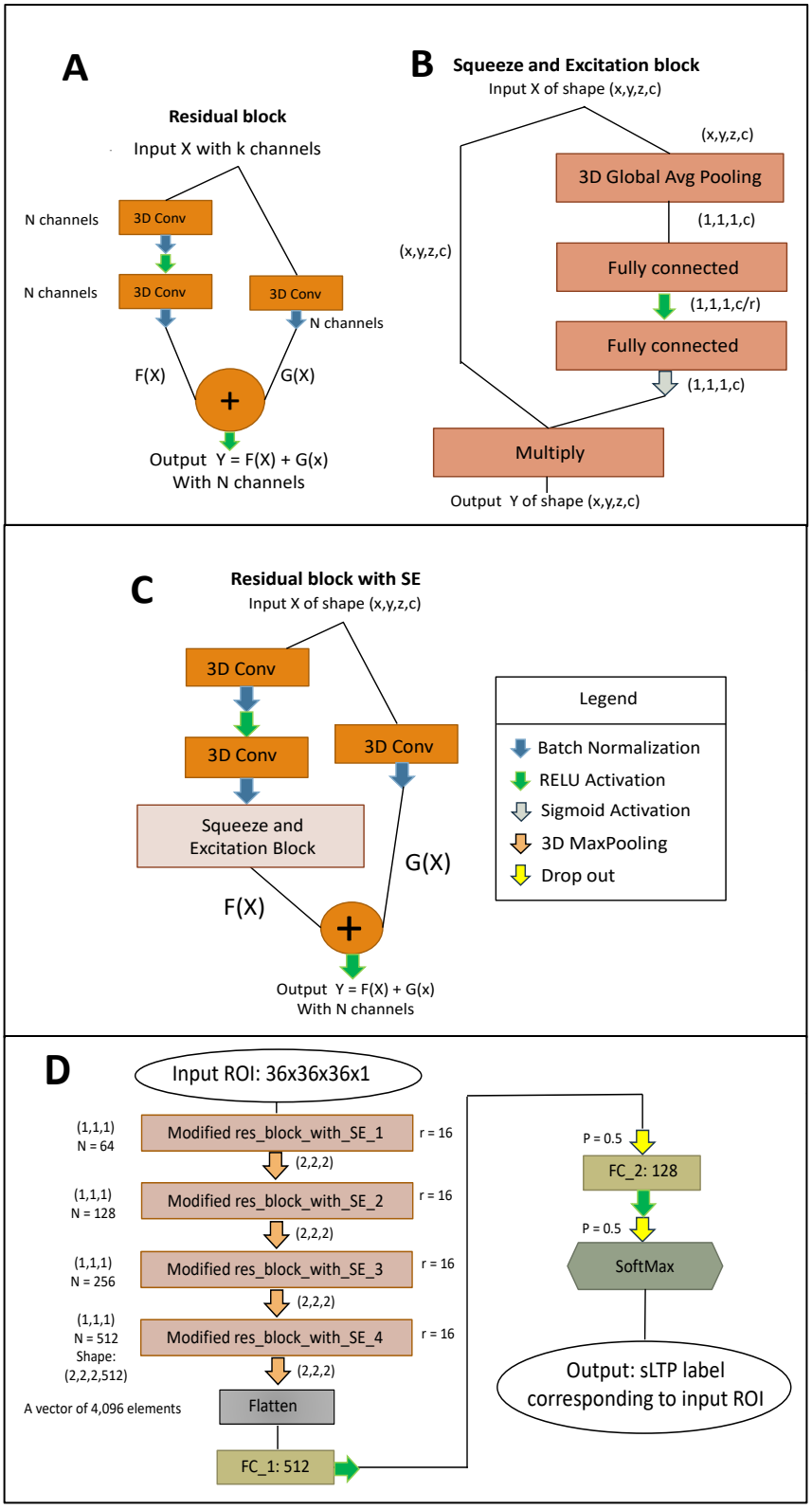

Figure 5.2: A) The architecture of the residual block used in this work, B) the architecture of an SE block, C) the modified residual block with SE, and D) the overall architecture of the proposed residual SE-CNN sLTP classifier.

5.3 Experiments and Results

5.3.1 Data and Preprocessing

In order to train, validate, and test our proposed residual SE-CNN sLTP classifier we used lung CT images from the SPIROMICS Visit 1 inspiratory cohort. This cohort includes 2,973 full-lung CT scans (2,400 participants with COPD, 600 without COPD, and 200 non-smoking controls) (Couper et al. 2014). Participants were 40-80 years old and imaging took place in six US field centers. A total of 10 scanner types were used for imaging. All participants underwent the CT scan at baseline and one-year follow-up. The lung CT scanning protocol in SPIROMICS is identical to that in MESA-Lung (detailed below) (Bild et al. 2002; Sieren et al. 2016). Images were reconstructed with an in-plane pixel resolution within the range [0.48, 0.98] mm, and a slice thickness of 0.625 or 0.75 mm (Couper et al. 2014). Specifically in this work, we used lung CTs from 2,922 distinct subjects from this cohort.

In addition to testing the model's performance on the SPIROMICS V1, we also tested the model on MESA-Lung exam 5 to assess its robustness to different cohorts and scanner types. It is worth mentioning that this dataset was used only for the purpose of testing the model, and no training procedure was applied using data from this cohort. MESA-Lung is a subset of the MESA study. The data in MESA-Lung exam 5 includes 3,200 subjects on 64-slice helical scanners (Couper et al. 2014) in accordance with MESA-Lung/SPIROMICS inspiration protocol detailed in (Sieren et al. 2016). Participants were 45-84 years old and free of clinical cardiovascular disease (Couper et al. 2014). In total, six scanner types were used for CT imaging. Images were reconstructed with an in-plane resolution within the range [0.47, 0.92] mm, and a slice thickness of 0.625 or 0.75 mm. In our experiment, we used lung CT images from 2,524 distinct subjects in this cohort.

The APOLLO® software (VIDA Diagnostic, Iowa) was used to generate lung masks for each subject in both SPIROMICS and MESA-Lung exam 5. Using the method in (Yang et al. 2021), sLTP masks were generated to create the ground truth labels on the segmented lungs. We linearly normalized all scans (separately for SPIROMICS and MESA-Lung) according to the global distribution of their intensity values in order to have an intensity range of [0,1] for training our residual SE-CNN model.

5.3.2 Detailed ROI Sampling Procedure and Data Augmentation

SPIROMICS data from 2,922 subjects was used to randomly sample ROIs of size 36×36×36 voxels from all 10 sLTP classes. ROIs were selected according to the following criteria: 1) For each 3D sampled ROI belonging to sLTP class i (*i.e.*, the sLTP class of the voxel in the center of the ROI), the percentage of that sLTP class was at least 30% (*i.e.*, percent $\text{sLTP}_i \geq 30\%$), and 2) no more than 20% overlap was allowed between any two ROIs. Since the distribution of sampled ROIs in the SPIROMICS dataset is nonuniform, we balanced the distribution of sampled ROIs according to the least frequent class by randomly selecting ROIs (without replacement) from the nonuniform distribution. Doing so helps the model to avoid running into class imbalance issues and getting stuck in local minima during the training procedure. Overall, 86,400 3D ROI samples were sampled across all 10 sLTP classes. To form train, validation, and test sets, a 6:2:2 split ratio was used. Fig. 5.3 depicts the balanced distribution of the 3D ROIs in SPIROMICS. In order to further enhance the training process by providing the model with more samples per class, we used a data augmentation technique applied only to the training set. Specifically, we flipped (mirrored) all 3D ROIs in the training set along each axis (*i.e.*, x, y, and z) resulting in a training set that was

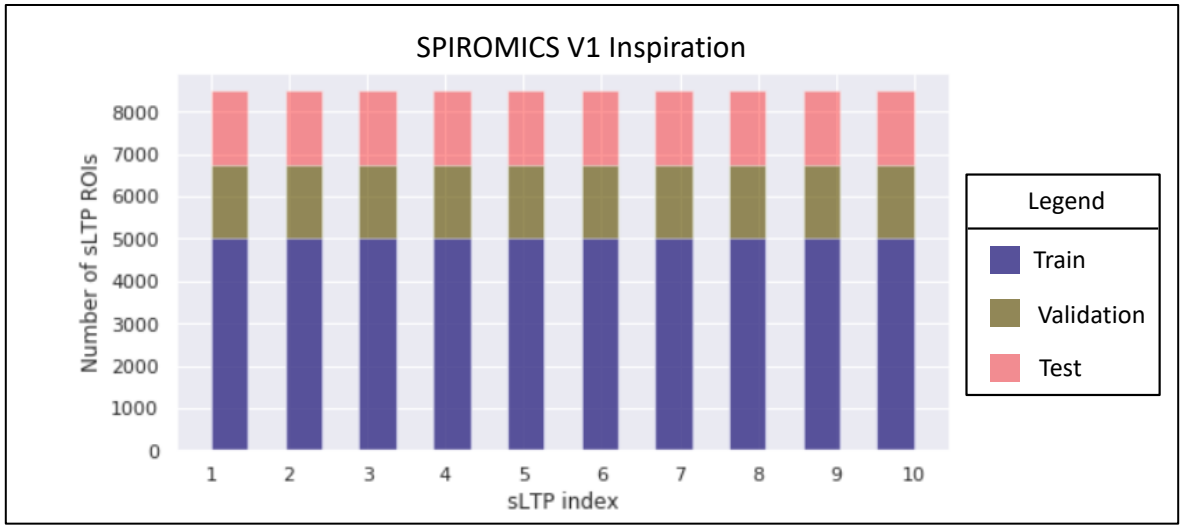

Figure 5.3: The balanced distribution of sampled ROIs in SPIROMICS divided into train, validation, and test sets using a split ratio of 6:2:2.

4 times larger (*i.e.*, the original ROI, plus the flipped versions along each axis). As a result, the model was provided with a total of 207,360 (20,736, per sLTP class) 3D ROIs in the training set.

We also used the MESA-Lung exam 5 dataset, in addition to the SPIROMICS dataset, to evaluate the robustness of our model across different cohorts. A total of 26,636 3D ROIs from 2,524 distinct subjects in MESA-Lung exam 5 were sampled based on the same criteria used in SPIROMICS (no data augmentation procedure was used). However, in order to evaluate the performance of the model according to a fair estimate of the true distribution of sLTPs in this cohort, this time we did not carry out any balancing procedure and tested the model against the actual, nonuniform sampled sLTP distribution for MESA-Lung exam 5 shown in Fig. 5.4.

5.3.3 Hyperparameters, Training History, and Performance on sLTP Classification

Fig. 5.2d shows the architecture of our model. As illustrated in Fig. 2, kernels of size [1, 1, 1] with no padding were used in all 3D convolutional layers to capture fine details within each sLTP class. For 3D convolutional layers, the number of kernels started at 64 and doubled after each residual block with SE to reach 512 by the 4th block. All four residual blocks with SE had a

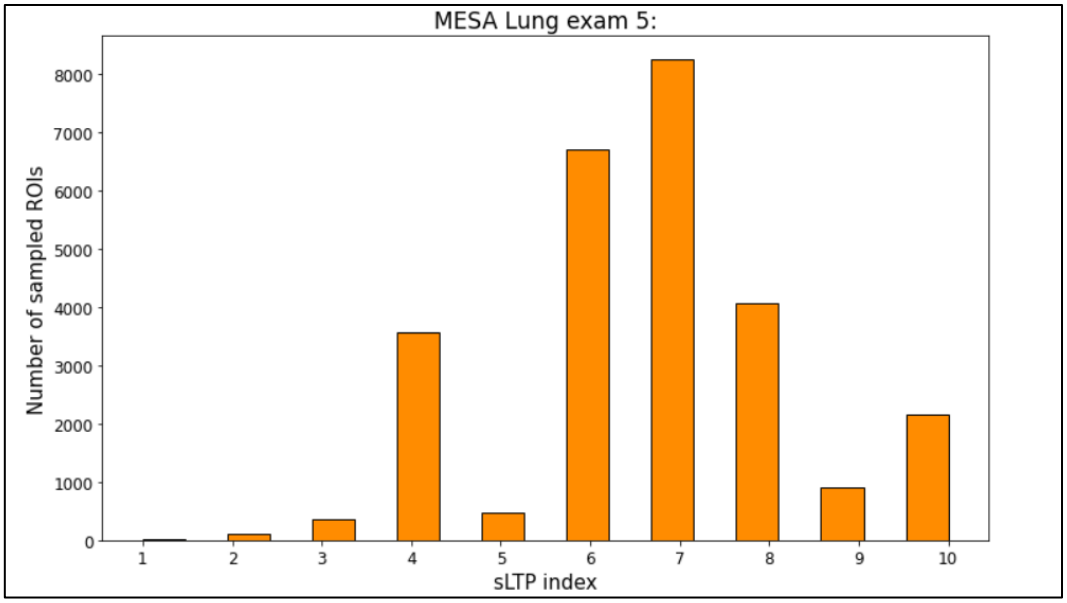

Figure 5.4: The distribution of sampled ROIs in MESA-Lung for testing the model's sLTP classification performance on a separate cohort.

fixed channel reduction factor *r* of 16. For all 3D Maxpooling layers the kernels were of size [2,2,2] without padding, thereby halving the feature maps in all directions. Furthermore, the probability of both Dropout units was set to 0.5 to minimize the overfitting effects associated with the two fully connected layers.

The residual SE-CNN sLTP classifier was implemented in Python using Keras with TensorFlow as backend on an Nvidia 2080 ti RTX (Titan RTX) GPU with 24GB memory. The model was trained for 350 epochs with a batch size of 32 to minimize 10 class Cross-entropy loss function. In order to minimize the cost function, stochastic gradient descent (SGD) with Nesterov optimizer (Sutskever et al. 2013) was used (initial learning rate = 0.0001, Momentum = 0.6, and decay rate = 10-6). Although the ADAM optimizer (Kingma and Ba 2014) has proven to be highly successful in training deep learning models substantially faster than SGD, it suffers from generalizability issues when compared to SGD (Zhou et al. 2020). Our model was initially trained using ADAM, but it did not generalize well to the test set. In keeping with the claim made in (Zhou et al. 2020) about SGD's generalizability power, we selected SGD in this study and used

momentum to speed up the training process. In Fig. 5.5, we present the training history of our model (the loss and accuracy as functions of epochs). We selected the model which provided the highest accuracy in our validation set as the most accurate model. Based on this, at epoch 336, the best model was found yielding 95.99% accuracy for training, 94.16% accuracy for validation, and 94.57% accuracy for the test set. Fig. 5.6 illustrates the confusion matrix for SPIROMICS V1's test set. According to Fig. 5.6, there are no significant misclassifications for any of the classes, and the matrix shows a diagonal pattern. Additionally, the average inference time of our model per ROI was 1.49ms.

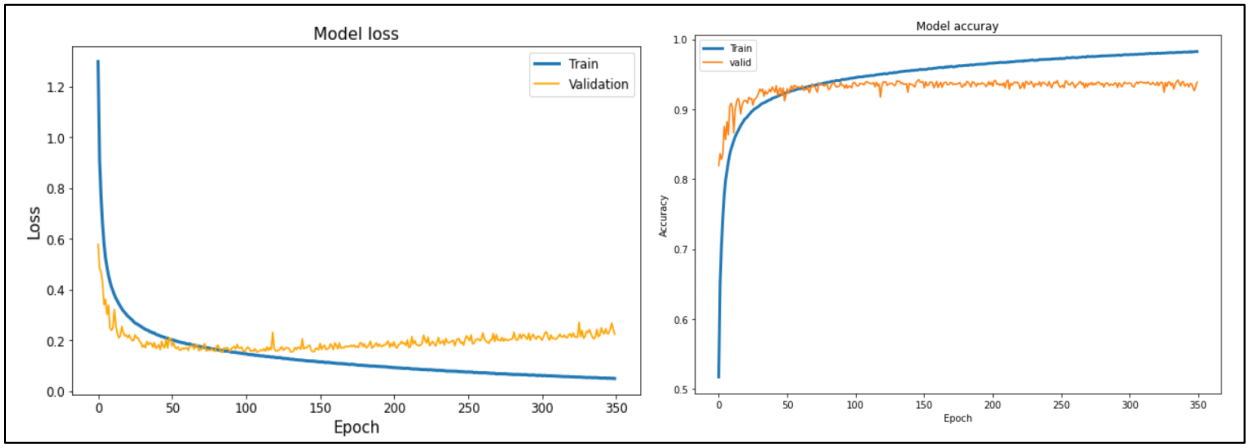

Figure 5.5: The training history of the residual SE-CNN sLTP classifier on SPIRMOICS V1. Best model is achieved at epoch 336.

As discussed earlier, our model was tested against the MESA-Lung exam 5 dataset in order to assess its robustness against a different cohort. The model was tested on all ROIs sampled from this cohort (distribution of the sampled ROIs is shown in Fig. 5.4). As a result, MESA-Lung exam 5 sLTP classification average accuracy was 93.85% (only 0.72% drop from SPIROMICS). Fig. 5.7, displays the confusion matrix for this analysis. The matrix shows a diagonal behavior with no per class classification accuracy below 90%, which indicates the robustness of our proposed sLTP classifier with respect to change in cohorts.

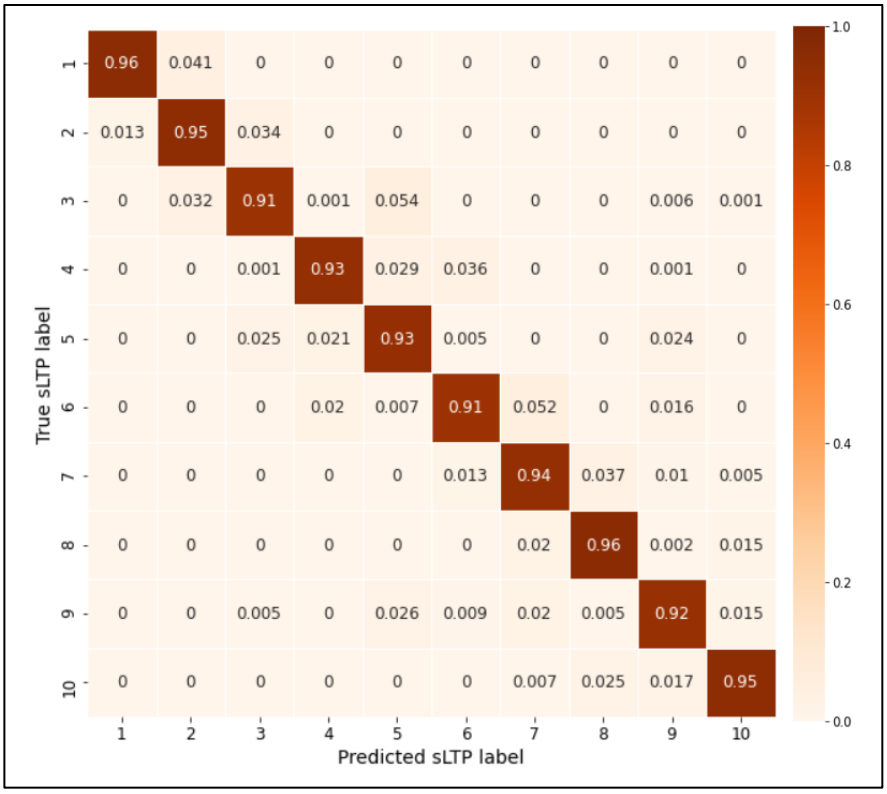

Figure 5.6: sLTP classification confusion matrix on SPIROMICS V1's test set.

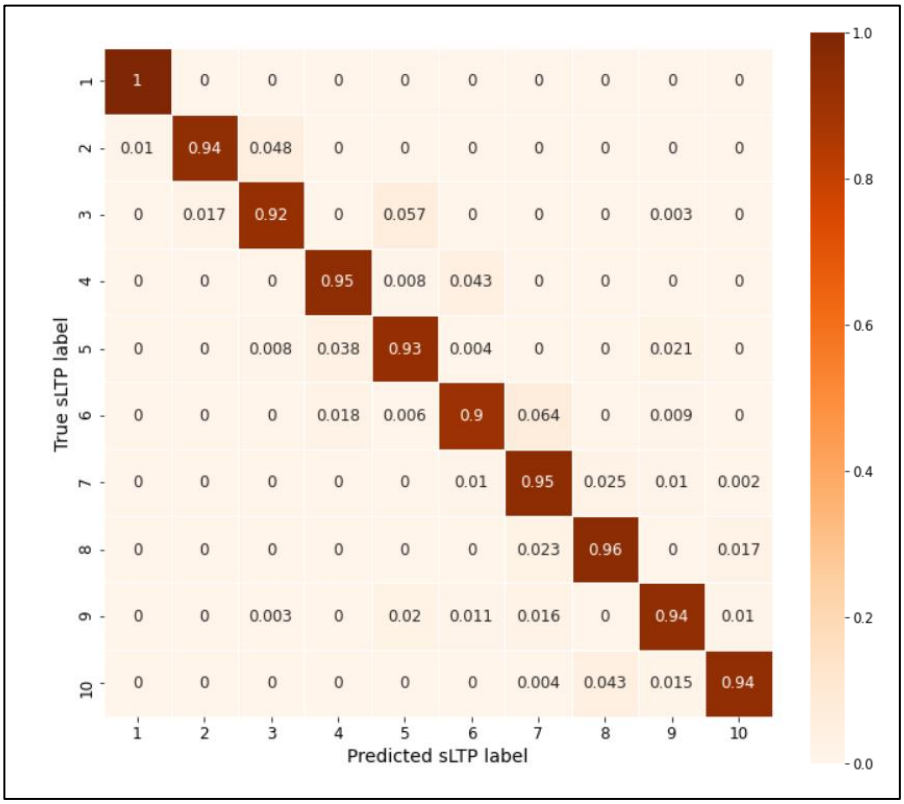

Figure 5.7: sLTP classification confusion matrix on the MESA-Lung exam 5 cohort.

5.3.4 Performance on CTES Classification

As noted previously, visual inspection and statistical analysis of sLTPs indicate that they can be divided into six CTES classes with clinical relevance (paper under review). The clustering of sLTPs is as follows: sLTPs 1 and 2 are grouped into CTES 1 named the vanishing lung, sLTPs 3, 5, and 9 are merged into CTES 2 named apical, sLTPs 4 and 6 are clustered into CTES 3 named diffuse, sLTP 7 forms CTES 4 named senile, sLTP 8 forms CTES 5 named obstructive Combined Pulmonary Fibrosis/Emphysema (CPFE), and sLTP 10 forms CTES 6 named restrictive CPFE (paper under review). Below we demonstrate how our model performs on SPIROMICS V1's test set as well as the MESA-Lung Exam 5 cohort with respect to CTES classification.

The confusion matrix for the 6 class CTES classification on SPIROMICS's test set is shown in Fig. 5.8. Overall, the CTES classification accuracy of the model is 96.02%. This matrix shows a diagonal distribution with minimal misclassification. Similar to the CTES confusion matrix shown in Fig. 5.8, MESA-Lung exam 5 also exhibits diagonal patterning as illustrated in Fig. 5.9. Based on the results of this cohort, the overall accuracy for 6 class CTES classification is 95.10% (only a 0.92% drop compared to SPIROMICS), demonstrating how robust our model is across different cohorts.

5.3.5 Performance Across Individual Scanner Types

The sLTP and CTES classification overall accuracies were quantified for each of the scanner types (separately for SPIROMICS and MESA-Lung) to provide evidence that the model can learn scanner-invariant features for sLTP classification. We present the results pertaining to the SPIROMICS test set and MESA-Lung in Tables 5.1 and 5.2, respectively. Almost all the scanner types performed similarly, indicating that the model is learning scanner-invariant features

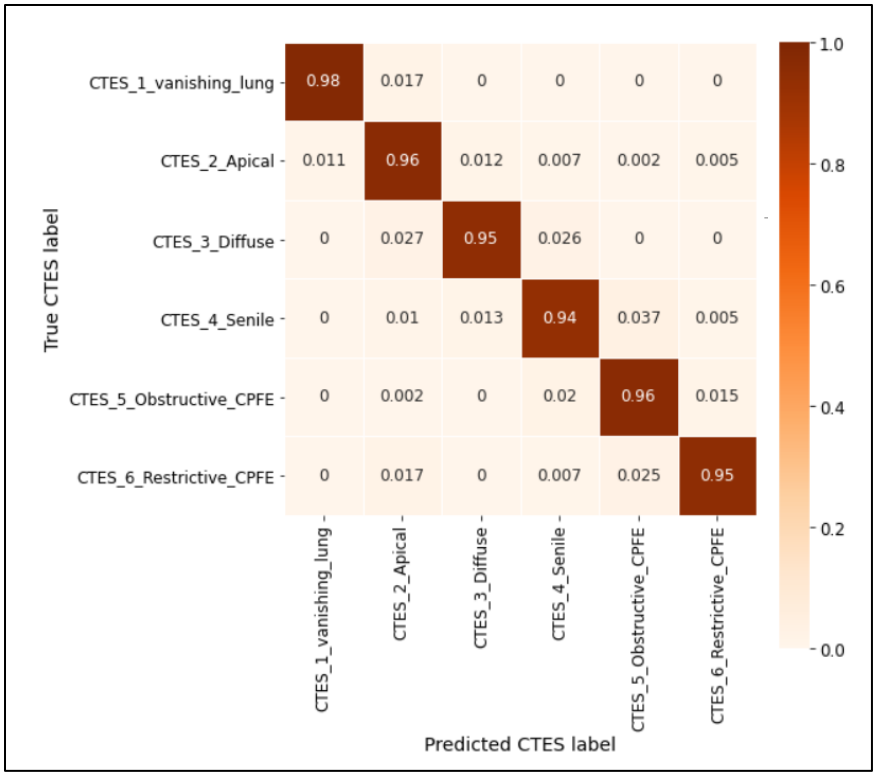

Figure 5.8: six-class CTES classification confusion matrix for SPIROMICS's test set.

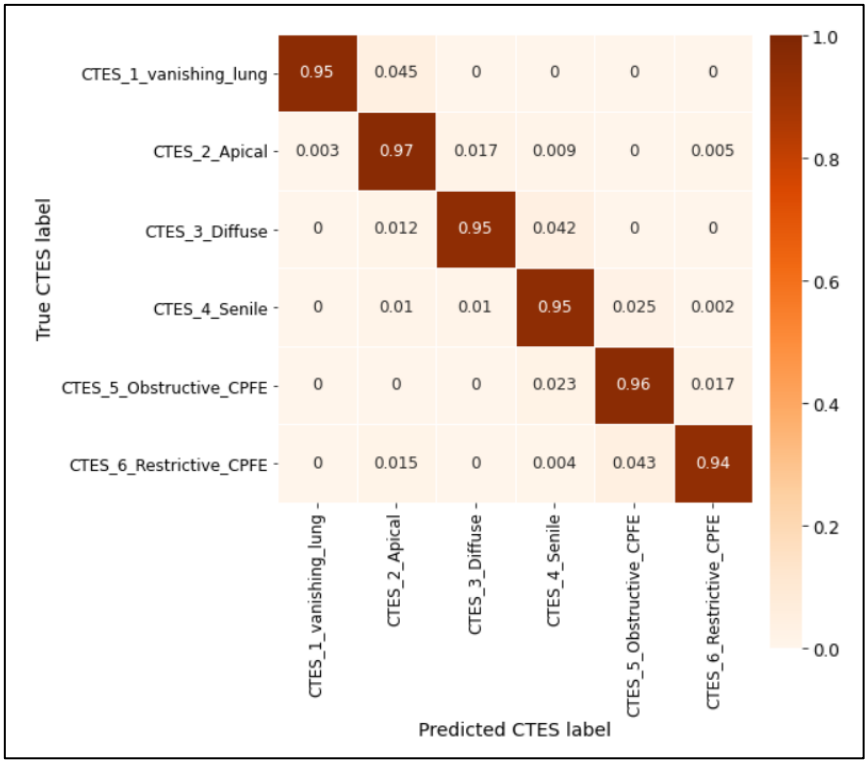

Figure 5.9: six-class CTES classification confusion matrix for MESA-Lung exam 5.

from sLTP data. Note that not all scanner types include every class of sLTPs/CTES. As some sLTP

classes are very rare in the MESA-Lung exam 5 cohort, we sampled twice as many ROIs in section 5.3.2 for this cohort in order to be able to report credible accuracies for different scanner types. Particularly, we sampled 30,900 additional ROIs from the MESA-Lung cohort in addition to the ones sampled in section 5.3.2., resulting in 57,536 ROIs for this analysis.

5.4 Discussion and Conclusion

As depicted in Figs. 5.6, 5.7, 5.8, and 5.9 for SPIROMICS and MESA-Lung cohorts, there is no significant misclassification in sLTP and CTES by the model. We demonstrated by training the model on the SPIROMICS cohort and testing it against an unseen cohort (*i.e.,* the MESA-Lung) that it is capable of learning cohort-invariant features for emphysema subtyping. Furthermore, our model learns scanner-invariant features from the data that are informative for sLTP and CTES classifiers based on our per-scanner accuracy results summarized in Table 5.1 and Table 5.2. However, SPIROMICS's and MESA-Lung's Definition and Discovery CT 750 HD scanners, respectively, perform less accurately than the other scanners in their respective cohort. This issue likely originates from the small number of ROIs that exist and have been sampled for these two scanners (*i.e.,* 11 ROIs for the Definition in SPIROMICS and 38 ROIs for the Discovery CT 750 HD in MESA-Lung). Despite these differences, the accuracies for all scanner types are comparable within in each cohort, indicating the model's ability to learn meaningful features for sLTP/CTES classification irrespective of scanner types. Our cohort and scanner-invariant model, in combination with the relatively fast ROI classification performance (average of 1.49 ms/ROI) makes our framework a quick and robust end-to-end classifier for emphysema learning and labeling.

Table 5.1: Scanner-specific sLTP and CTES classification accuracy in SPITOMICS's test set

	SPIROMICS V1 inspiration cohort scanner types									
	Definition	Definition AS+	Discovery CT750 HD	Discovery STE	LightSpeed VCT	SOMATOM Definition	SOMATOM Definition AS	SOMATOM Definition AS+	SOMATOM Definition Flash	Sensation 64
sLTP classification accuracy	90.90% (11 ROIs) (2 Subjects)	93.93% (825 ROIs) (86 Subjects)	93.61% (4380 ROIs) (486 Subjects)	93.83% (1507 ROIs) (154 Subjects)	93.77% (5650 ROIs) (655 Subjects)	92.92% (1525 ROIs) (198 Subjects)	95.07% (406 ROIs) (32 Subjects)	92.02% (129 ROIs) (208 Subjects)	94.35% (902 ROIs) (157 Subjects)	92.22% (784 ROIs) (55 Subjects)
Existing sLTP classes	4, 5, 6, 7	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
CTES classification accuracy	90.90% (11 ROIs) (2 Subjects)	97.21% (825 ROIs) (86 Subjects)	95.59% (4380 ROIs) (486 Subjects)	95.69% (1507 ROIs) (154 Subjects)	96.00% (5650 ROIs) (655 Subjects)	96.33% (1525 ROIs) (198 Subjects)	98.03% (406 ROIs) (32 Subjects)	96.12% (129 ROIs) (208 Subjects)	96.90% (902 ROIs) (157 Subjects)	95.28% (784 ROIs) (55 Subjects)
Existing CTES classes	2, 3, 4	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6

Table 5.2: Scanner-specific sLTP and CTES classification accuracy in MESA-Lung

	MESA Lung exam 5 cohort scanner types					
	Definition	Discovery CT750 HD	Discovery STE	LightSpeed VCT	Sensation 64	Sensation Cardiac 64
sLTP classification accuracy	92.94% (1614 ROIs) (100 Subjects)	89.47% (38 ROIs) (3 Subjects)	93.41% (8482 ROIs) (372 Subjects)	93.89% (11,628 ROIs) (580 Subjects)	92.52% (22,750 ROIs) (945 Subjects)	92.80% (13,024 ROIs) (524 Subjects)
Existing sLTP classes	2, 3, 4, 5, 6, 7, 8, 9, 10	4, 6, 7, 8, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
CTES classification accuracy	95.35% (1614 ROIs) (100 Subjects)	89.47% (38 ROIs) (3 Subjects)	94.95% (8482 ROIs) (372 Subjects)	94.64% (11,628 ROIs) (580 Subjects)	94.92% (22,750 ROIs) (945 Subjects)	95.27% (13,024 ROIs) (524 Subjects)
Existing CTES classes	1, 2, 3, 4, 5, 6	3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6	1, 2, 3, 4, 5, 6

Chapter 6: Investigation and Characterization of Ipsilateral and Contralateral Negative BOLD Response Patterns in Human Brain fMRI Scans[1]

6.1 Introduction

fMRI stands for functional magnetic resonance imaging. It is a non-invasive imaging tool used to study human brain activity *in-vivo* via endogenous blood oxygenation level dependent (BOLD) contrast (Ogawa et al. 1992). When a targeted stimulus evokes neuronal activity in part(s) of the brain, the MR signal in the activated area(s) usually changes in comparison to its pre-stimulus state. This change in MR signal is referred to as the positive BOLD response (PBR), and it has been characterized as the hemodynamic response function (HRF) of the BOLD signal. Aside from the PBR, fMRI responses in the opposite direction have also been observed in regions near and remote to the PBRs. The inverse direction BOLD responses are referred to as negative BOLD responses (NBRs).

PBR hemodynamics and their underlying mechanisms have already been extensively studied, but NBR hemodynamics and their underlying mechanisms have yet to be fully explored (Logothetis 2008; Liu et al. 2011). Studies have reported observing NBRs following sensory cortical unilateral stimulation in proximity to the PBR (Shmuel et al. 2002; Shmuel et al. 2006), as well as in ipsilateral regions (Shmuel et al. 2003; Stefanovic, Warnking, and Pike 2004; Smith A.T. et al. 2004; Kastrup et al. 2008; Mullinger et al. 2014). However, it is unknown how these two types of NBRs emerge and what their properties are. It has been hypothesized that NBRs may

[1] Disclaimer: this work was done under supervision of Dr. Qolamreza "Ray" Razlighi at Quantitative Neuroimaging Laboratory of Columbia University Medical Center between 2018 to 2019. Hengda He, and Dr. Amir Shmuel are also coauthors on the draft of the manuscript of this work that, at the time of this writing, is being prepared for re-submission to a journal of interest. Parts of this chapter are taken from the draft of the manuscript.

result from the following mechanisms: 1) Inhibition of neuronal activity at the local level (Smith A.T. et al. 2000; Smith A.T. et al. 2004; Shmuel et al. 2002; Shmuel et al. 2006; Stefanovic et al. 2004), 2) Reduced cerebral blood flow (CBF) in regions close to PBR, possibly as a result of sharing upstream arterial supply (known as blood steal), irrespective of changes in neuronal response (Shmuel et al. 2002; Harel et al. 2002; Hu and Huang 2015), 3) Active reduction in CBF (known as blood sharing) caused by neural control signals leading muscles to contract around the arteries supplying the NBR regions (Smith A.T. et al. 2004), and 4) Venous back-pressure from limited drainage capacity of venous compartments as a result of increased CBF in the PBR region (Shmuel et al. 2006; Boas et al. 2008; Goense, Merkle, and Logothetis 2012). These mechanisms, however, cannot account for simultaneous observation of NBRs adjacent to PBR and in the hemisphere opposite to the one receiving the stimulation.

During unilateral visual hemifield stimulation, two types of NBR can be detected concurrently: contralateral NBR (cNBR) from the proximal regions of the PBR, and ipsilateral NBR (iNBR) from the other hemisphere. Therefore, the question arises as to what are the temporal and spatial influences between these two NBRs, as well as whether their relationship supports or refutes the existing hypotheses. To investigate this, we measured PBR and both NBRs elicited in human visual cortex following unilateral visual stimulation. We first determined whether the magnitudes of the two NBRs have a linear relationship with the duration of the stimulus since linearity of the BOLD response is crucial for many statistical analyses performed in fMRI. Our next step was to extract the HRFs of both NBRs to determine whether their magnitudes and dynamics differed. Additionally, we explored the correlation between the subject-wise expression of both NBRs and the PBR, their interhemispheric spatial similarity, and their behavioral relationship. Lastly, we studied whether attention affected the two NBRs. Our observations suggest

that the two NBRs may share a common neural and/or vascular mechanism, which is controlled by a brain structure likely located deep within the cerebral cortex controlling cerebral blood flow distribution.

6.2 Method

6.2.1 Overview

In order to investigate the characteristics of cNBR and iNBR, we first designed a specific fMRI experiment to elicit these two responses in human visual cortex. After data collection, we used fMRI preprocessing steps to extract the signals of interest. Following the extraction of all 3 BOLD responses (*i.e.,* PBR, cNBR, and iNBR), we carefully investigated their properties through conducting a series of analyses which are detailed in the following sections.

6.2.2 Data Collection and fMRI Preprocessing Steps

In this study, 42 healthy adults (mean age ± SD = 40.05 ± 20.56, female/male = 26/16, right-handed) were recruited. Random mailings were used to recruit candidates within a 50-mile radius of Columbia University Irving Medical Center. All participants were given compensation for the time they spent participating in this study. Written informed consent was obtained prior to the scanning sessions. Columbia University's institutional review board approved our study design and recruitment process.

Our fMRI experimental task design utilized an event-related approach. Task components included two types of ongoing stimuli: 1) A flashing checkerboard with maximum contrast (visual stimulus) displayed either on the left or right side of the screen, and 2) A right- or left-sided alternating tone (auditory stimulus) paradigm played on the ear. We administered both sensory stimuli with random onsets and durations (uniform distribution, durations ranging from 1 to 5 seconds). While visual and audio stimulation overlaps were permitted, bilateral exhibition of the

same sensory stimulus (left-visual with right-visual or left-auditory with right-auditory) was disallowed. Data collection took place in 2 separate runs (happening on the same day). During the first run, participants were instructed to focus on only one sensory modality (e.g., visual or tonal) while neglecting the other. In the second run, we instructed them to pay attention to the other modality which they ignored in the first run. There were a total of 120 events presented in each scan session, 60 of which were visual and 60 of them were auditory. Furthermore, half of the events in each modality were confined to one side (left or right) and were presented at inter-stimulus intervals ranging from 1 to 16 seconds, selected from a uniform distribution. As soon as the attended stimulus stopped, subjects were asked to press a button twice with their right/left index fingers (contingent on the right/left polarity of the attended stimulus) in order to track their attendance. Additionally, all subjects were instructed to continuously maintain their gaze on a tiny green fixation point in the middle of the screen throughout the scan. The fixation point's color was changed to red anytime the subject's response was incorrect or out of time (over 3 seconds). Using an eye-tracking system, the position of the eyes was continuously monitored to determine eye fixation. Prior to scanning, subjects practiced outside of the scanner to learn and perform the task correctly on short training trials. A diagram of an example of this fMRI experiment is shown in Fig. 6.1.

A 3 Tesla Siemens Magnetom Prisma MRI scanner was used to image all participants. For capturing fMRI signals, the $T2^{*}$-weighted echo-planar imaging (EPI) sequence was used with interleaved slice acquisition. The fMRI acquisition parameters are summarized in Table 6.1. During the experiment, subjects wore noise-isolating MR safe earbuds to listen to the auditory stimuli while placed supine inside the magnet bore. An area placed in the center of each hemifield contained the flashing checkerboard visual stimuli. A screen with back-projection was used to

display the visual stimuli in the scanner. The task-based fMRI (tb-fMRI) scans took approximately 8 minutes (equal to 480 volumes). Participants also underwent a resting-state fMRI (rs-fMRI) scan with the same parameters and duration. For registration of fMRI volumes, a T1-weighted image was also obtained for each participant. Details of this acquisition are summarized in Table 6.1.

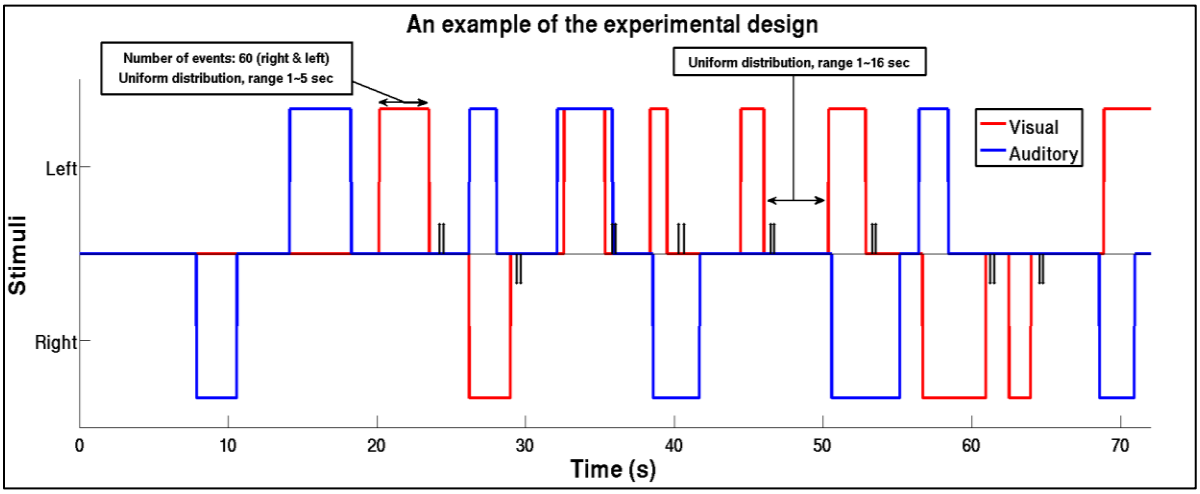

Figure 6.1: An example of the fMRI task design. In the blue line are the timings for the auditory stimuli, while in the red line are the timings for the visual stimuli, and in the black line are the responses from the subject (*i.e.*, when they hit the button). Subjects were asked to ignore the auditory stimulus for this sample demonstration and focus on the visual stimulus. An indication of this is the pattern of responses, where the button is pressed twice by the subject as soon as the visual stimulus has ended.

Table 6.1: Acquisition parameters for all T2*-weighted and T1-wieghed images used in this work

Parameters	T2*-weighted	T1-weighted
TR/TE (ms)	1000/30	2300/2.32
Flip angle (degrees)	62	8
Field of view (mm)	200 × 200	256 × 256
Matrix size	100 × 100	256 × 256
Voxel size (mm)	2 × 2 × 2	1 × 1 × 1
Number of slices	64 axial slices[1]	196 sagittal slices
Bandwidth (Hz/px)	1852	-
Multiband factor	4	-

[1] With alternating phase encoding directions of anterior-posterior and posterior-anterior

FSL V5.0.7 (Smith S.M. et al. 2004) and in-house software packages were employed to process the fMRI data. The raw fMRI time-series were corrected for slice acquisition delays via slice timing correction techniques (Parker, Liu, and Razlighi 2017; Parker and Razlighi 2019). Meanwhile, raw fMRI scans (uncorrected for slice timing) were used to estimate motion parameters via registrations of all volumes to the first one. Furthermore, using the topup algorithm (Andersson, Skare, and Ashburner 2003) included in FSL, the compensatory fields for adjusting susceptibility-induced geometric deformations were calculated from the first volumes of the raw fMRI scans with opposite phase encoding direction. Following this, the motion parameters along with the estimated compensatory field were applied together to the slice timing corrected fMRI time-series to clear the data from such artifacts. The preprocessed fMRI data were modeled with 6 predictors (aside from the linearity analysis) for all 6 stimulation categories (*i.e.*, left and right visual, audio, and motor responses). The canonical double-gamma HRF (Friston et al. 1998) was used to obtain the 6 predictors through convolution with their respective stimulation. An in-house Python software package was used to perform first-level analysis via multiple regression. Finally, the T1-weighted scans were processed by Freesurfer (Fischl et al. 2002; Fischl et al. 2004).

6.2.3 Linearity Analysis

Here, we examine the linearity of the two NBRs in relation to the change in stimulation duration. In terms of duration, stimuli are grouped into four distinct categories (category 1 includes stimuli with durations of 1 to 2 seconds, while categories 2, 3, and 4 include stimuli with durations of 2 to 3, 3 to 4, and 4 to 5 seconds, respectively). Our standardization procedure involves making the fMRI time series have zero mean and dividing them by their standard deviation. Voxel selection is carried out using a bootstrapping technique in which each trial is randomly split in half for voxel selection (training) and BOLD response extraction (testing). Such a method of voxel

selection prevents the analysis from having any bias in favor of voxels showing random linear properties by luck. During each epoch, general linear modeling (GLM) analysis is employed on the training boxcars to identify the significantly responsive voxels (t-test, $|t| > 3$). The selected voxels are then used in the testing trials to examine the linearity. To ensure a unit area under the curve in each trial's boxcar, the amplitude of each trial in the testing set is normalized according to its duration. Depending on the categorized stimulus durations, each trial is then classified into one of the four different regressors. Using each of the four regressors, four separate GLM analyses are then conducted to estimate the BOLD response magnitude for each of the duration categories. During the booting process, 500 epochs are repeated. To calculate the corresponding BOLD response amplitude, the resultant cNBR and iNBR point estimates are averaged (separately) across all epochs and stimulation lateralization (i.e., right and left sides). A plot is then drawn showing the amplitudes of subject-wise mean point estimates corresponding to each duration class against their four duration categories. Lastly, we perform a least-squares line regression analysis and a t-test of the slope to determine if there is a significant linear relationship between the amplitudes and stimulus duration. This method is also used as a sanity check to determine the linearity of the PBR.

6.2.4 Generating ROIs

In this work, our group-level analysis is conducted in the subjects' native space (unless otherwise noted), not in standard space (e.g., brain atlas). In comparison with the regular group-level analysis in the standard space, native space analysis has the advantage of being more precise at localizing ROIs in each subject. In this manner, two ROIs are created per hemisphere for each subject. For each hemisphere, the first ROI is constructed by selecting the voxels where PBR is statistically significant using the z-statistical map derived from the first-level analysis ($z > 4$). To

locate the neighboring voxels required for identifying the cNBR, this ROI is then dilated using a spherical kernel with a radius of 10 mm. To detect voxels with iNBR, we use the union of two masks on the opposite hemisphere as the searching ROI. The two generated ROIs are then further refined by eliminating any intersection with the inferior parietal and precuneus regions. This refinement was made in light of the fact that the NBRs elicited by these two regions might be part of the responses associated with the default mode network (DMN), which may possess completely different properties. Lastly, each bilateral ROI is restricted to fall within each hemisphere of the brain to avoid any potential overlap or signal leakage.

6.2.5 Hemodynamic Response Functions Extraction

For the purpose of investigating the two HRFs' shapes, voxel selection cannot be conducted using the canonical double Gamma HRF. The reason for this is to prevent bias towards a particular HRF shape. As a result, we select voxels for this experiment using the FMRIB's linear optimal basis set (FLOBS) method (Woolrich, Behrens, and Smith 2004) which does not suffer from bias issues. The criterion for choosing the significantly responsive voxels within the ROIs extracted in the previous section was set to F statistics of $|F|>6$. Based on the estimated beta values in the first-level FLOBS GLM analysis, an initial response function is computed for each voxel by averaging over the FLOBS basis functions. To identify the direction (i.e., the sign) of the extracted BOLD response, we utilize the sign of the measured area under the curve of the first 10 seconds of the response function. Then, the FIR deconvolution method (Goutte, Nielsen, Hansen 2000) is applied to extract the impulse response function from the time series of the selected voxels. Using the explained procedure, all voxel-level iNBR and cNBR HRFs corresponding to each subject, are extracted for both attended and unattended stimuli. Finally, the HRFs are averaged for each subject

with respect to lateralization of the visual stimuli to obtain the average subject-level HRFs. The same exact process is utilized to extract the PBR HRF.

6.2.6 Subject-wise Expressions for BOLD Responses

In this analysis, we intend to examine the relationship between subject-wise expressions of BOLD responses, which may provide clues to the underlying cause of NBRs. Expressions for each subject are calculated using point estimates from significantly affected voxels. Pearson correlation coefficient (PCC) is used separately for the attended and unattended cases for each pair of BOLD amplitudes (*i.e.,* PBR vs cNBR, PBR vs iNBR, and cNBR vs iNBR) to assess their relationship. Lastly, the significance of the difference in correlation values between pairs of BOLD responses is tested using the statistical test of the difference between two correlation coefficients.

6.2.7 Analysis of Interhemispheric Spatial Similarities

This analysis aims to explore the relationship between the spatial patterns of BOLD responses of the opposite hemispheres (*i.e.,* iNBR vs PBR, and iNBR vs cNBR). Dice similarity coefficient (DSC) is used as the metric for evaluating the strength of spatial similarity. Bilateral ROIs are mapped into MNI standard space (Evans et al. 1993) in order to measure the spatial pattern similarity across them. Following that, by flipping the mapped ROIs along the mid-sagittal plane and using the DSC score the strength of their spatial similarity is calculated. The student t-test is then used to assess the significance of the difference (if any) between the calculated DSC scores.

6.2.8 Behavioral Correlates of the BOLD Responses

As a performance measure, we use the subject-wise median response time to examine the relationship between the NBR and task performance. PCC was used to measure the potential

association (if any) between the task performance and corresponding expressions for PBR, cNBR, and iNBR.

6.2.9 Analyzing Attention Dependency of the BOLD Responses

The purpose of this analysis is to determine whether attention alters BOLD signals. We define attention dependency as the difference of the amplitudes and dynamics of the extracted HRFs between attended and unattended cases. We use the student t-test (with Bonferroni correction) to measure the significance of any difference (if there is one) between attended and unattended cases at each second following stimulus onset.

6.3 Experiments and Results

This section presents the results of our analyses. Independent of the attention case, all subjects showed sustained PBR, iNBR, and cNBR due to visual stimulation in visual areas of the Occipital lobe. To demonstrate this, an overview of a sample subject's PBR, cNBR, and iNBR in response to a unilateral display of attended visual stimuli on the right hemifield is shown in Fig. 6.2 using z-statistics heatmap superimposed on three orthogonal slices of the subject's T1-weighted image.

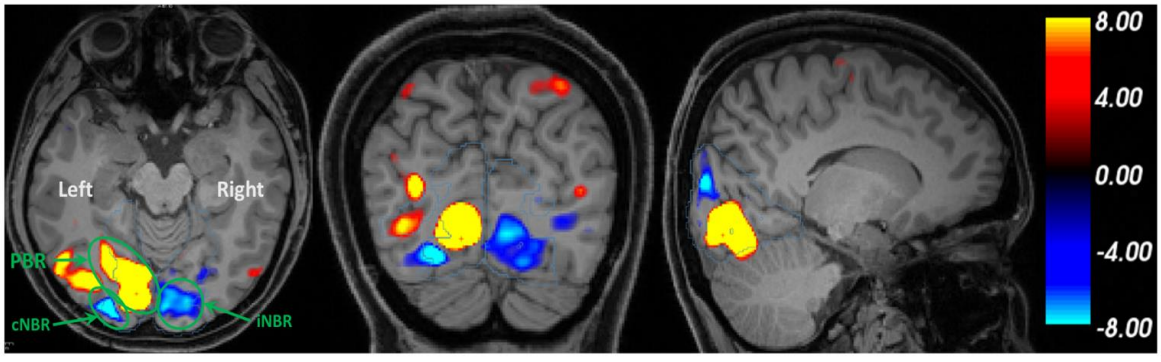

Figure 6.2: An example of the spatial patterns of the fMRI responses of a subject to the attended right hemifield visual stimulation. On the Axial slice (the left most image), statistically significant PBR, cNBR, and iNBR are shown using the green arrows and circles.

6.3.1 Stimulus Duration vs BOLD Magnitude Linearity

The averaged point estimates for cNBR, iNBR, and PBR are plotted as a function of stimulus duration in Fig. 6.3. As depicted in this figure, the amplitudes of both cNBR and iNBR are linearly related to stimulus duration. The significance of this linear relationship for both NBRs is quantified as: cNBR: $\beta = -0.298$, $p < 0.015$; and iNBR: $\beta = -0.184$, $p < 0.031$. In simple terms, the results of this analysis suggest that a longer stimulus can elicit a more negative (*i.e.*, larger amplitude) cNBR and iNBR in the stimulated regions of the brain. To confirm whether or not our approach of assessing linearity is a valid method, we also checked the linearity of PBR (which has been shown previously in the literature (Boynton et al. 1996)) with it. Using our method, the linearity of PBR with respect to the stimulus duration holds ($\beta = 0.810$, $p < 2.134 \times 10^{-4}$).

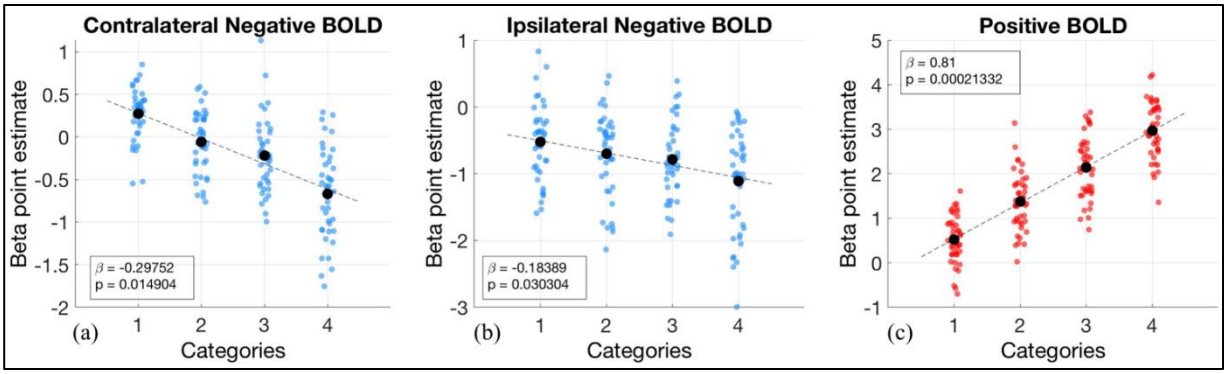

Figure 6.3: Results of the linearity analysis. For each duration category, the black dots represent the mean (over subjects) amplitude of the corresponding BOLD response. Regression lines are depicted with black dashed lines. Both iNBR and cNBR show statistically significant linear relationships with stimulus duration.

6.3.2 Relationship Between HRFs of cNBR and iNBR

In Fig. 6.4, we present the HRFs of all three BOLD responses for both attention cases. Irrespective of the attention case, the shape of cNBR and iNBR HRFs are fairly similar. As a measure of similarity, we examined whether there was any statistically significant difference in the amplitudes, as well as onset, peak, falling edge, and undershoot times of the two NBRs using the individual subject-level HRFs. No significant differences between the amplitudes of the two

NBRs were detected in both the attended (Δ = -1.34%, $p > 0.8975$) and the unattended (Δ = 2.30%, $p > 0.7444$) groups. Additionally, neither the attended (Δt = -0.1517 s, $p > 0.3382$) nor the unattended (Δt = -0.0031 s, $p > 0.9834$) cases exhibited any significant difference in the onset times of the two NBRs. Similar results were obtained with respect to the difference in the peak time of the two NBRs (attended Δt = -0.0093 s, $p > 0.9536$; unattended Δt = -0.0699 s, $p > 0.7915$). Moreover, neither of the NBRs' falling edge times were significantly different in the attended condition (Δt = -0.3666 s, $p > 0.5117$). In contrast, the unattended (Δt = -1.4657 s, $p < 0.0252$) case demonstrated a significant difference, but it failed to survive the multiple comparison correction. Lastly, in terms of undershoot time for both the attended (Δt = -0.5155 s, $p > 0.5520$) and unattended (Δt = -0.0707 s, $p > 0.9274$) cases, there was no statistically significant difference between the two NBRs. These results suggest that the two NBRs are likely to arise from the same mechanism. We also compared the iNBR and cNBR HRFs with the PBR HRF. Fig. 6.5 demonstrates the result of this comparison. The statistically significant differences are highlighted with asterisk symbols in the figure.

6.3.3 Association of Subject-Level Expressions of cNBR and iNBR

This experiment sought to determine whether cNBR and/or iNBR are driven by the PBR in the visual cortex. In light of this, we compared the magnitude of the relationships between PBR and the two NBRs, with the strength of the relationships between the two NBRs. The scatter plots in Fig. 6.6 depict the subject-wise mean expressions of iNBR versus cNBR, PBR versus iNBR, and PBR versus cNBR, for both attention cases separately. Even though the PBR and cNBR were elicited in regions close to each other on the same brain hemisphere, the cNBR showed a significantly greater correlation with the iNBR than with the PBR (attended case: $z = 3.1001$, $p <$

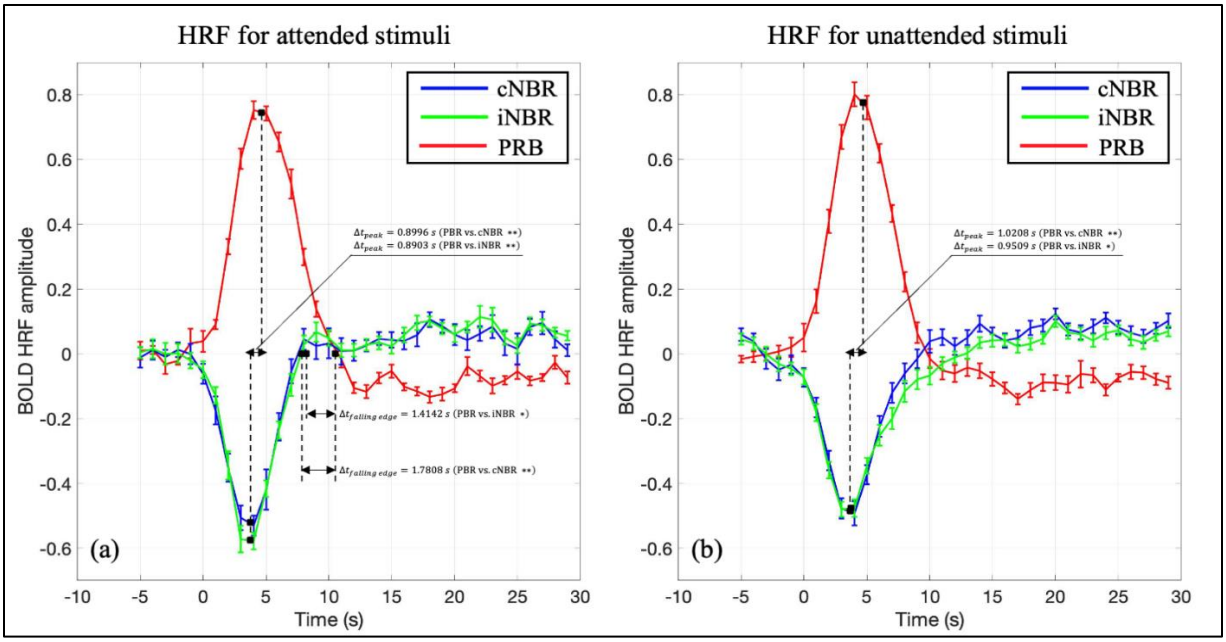

Figure 6.4: Extracted HRFs of the BOLD response for: a) the attended, and b) the unattended cases. The HRFs are adjusted by averaging their values for 5 seconds before the stimulus begins. HRF magnitudes are calculated using percent change. Error bars indicate the standard error of the mean. The iNBR and cNBR's HRFs are significantly different from one another (* $p < 0.05$; ** $p < 0.01$; Bonferroni correction).

0.0019, and unattended case: $z = 2.9969$, $p < 0.0027$). This suggests that PBR is less likely to directly cause the cNBR or iNBR.

6.3.4 Spatial Patterns Similarity Between Interhemispheric BOLD Responses

We examined the spatial similarity of the visual patterns of the evoked BOLD responses in the brain. Fig. 6.7 demonstrates the result of this analysis on inflated visual regions of the brain. The regions associated with iNBR and cNBR showed minimal spatial similarity quantified by the dice score (DSC = 0.0517 ± 0.0249; mean ± SD). This similarity was significantly lower ($p <$ 0.0163) than the spatial similarity between regions occupied by iNBR and PBR (DSC = 0.0668 ± 0.0312; mean ± SD) for the attended visual stimulation. This observation provides evidence that the structural connections between the two hemifields such as Callosal pathways (Bocci et al. 2014; Fabri et al. 2011) could not account for the tight coupling of the two NBRs demonstrated in the previous analysis.

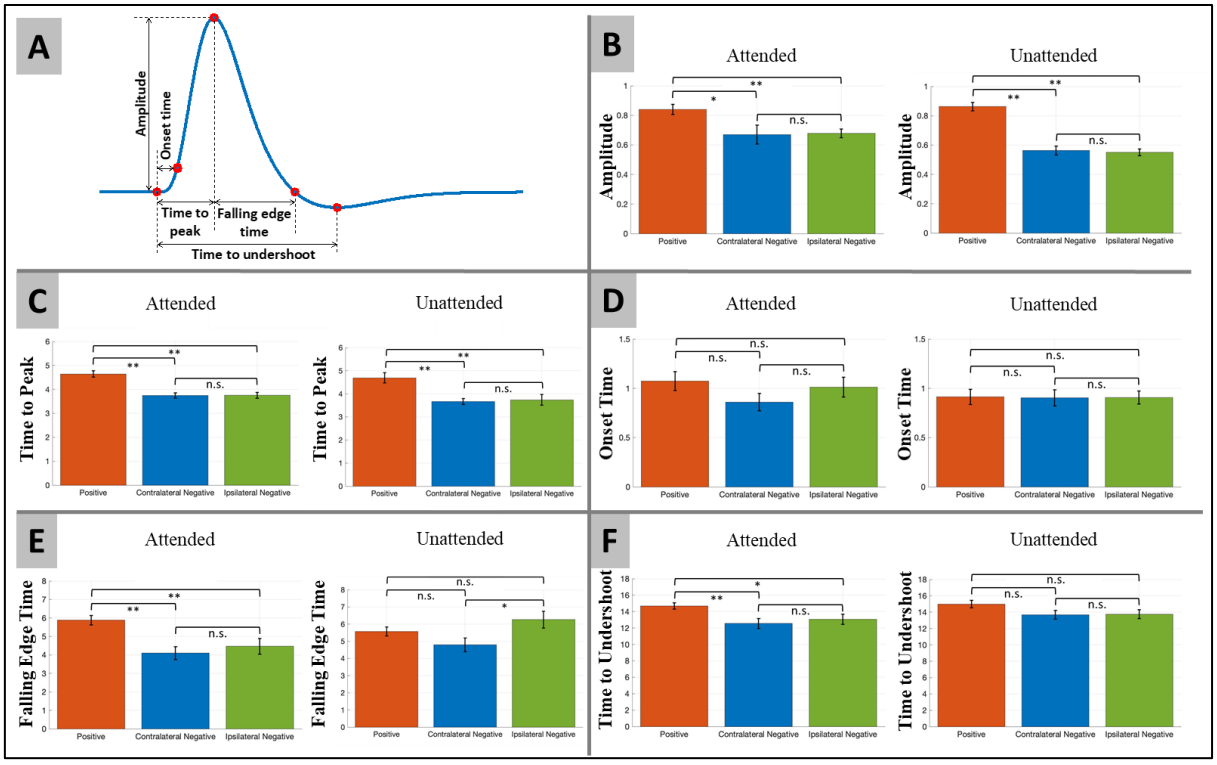

Figure 6.5: A) Definitions of amplitude, as well as peak, onset, falling edge, and undershoot times of the BOLD signals. B-F) Student's t-test of any statistically significant difference between the B) amplitude, C) peak time, D) onset time, E) falling edge time, and F) undershoot time of the two NBRs and PBR HRFs. Error bars indicate the standard error of the mean. Statistically significant differences are marked with asterisk symbols (* $p < 0.05$; ** $p < 0.01$; uncorrected).

6.3.5 Relationship with Task Performance

We investigated the relationship between NBRs and task performance, as mentioned previously. Fig. 6.8 illustrates the result of behavioral correlates of all 3 BOLD signals. As shown, despite a marginally significant relationship between PBR and performance ($r = -0.1799$, $p > 0.2541$), the cNBR ($r = 0.0469$, $p > 0.7678$) and the iNBR ($r = 0.0504$, $p > 0.7510$) did not show any statistically significant relationship. This suggests that neither of the two NBRs may directly contribute to task performance or get modulated by it.

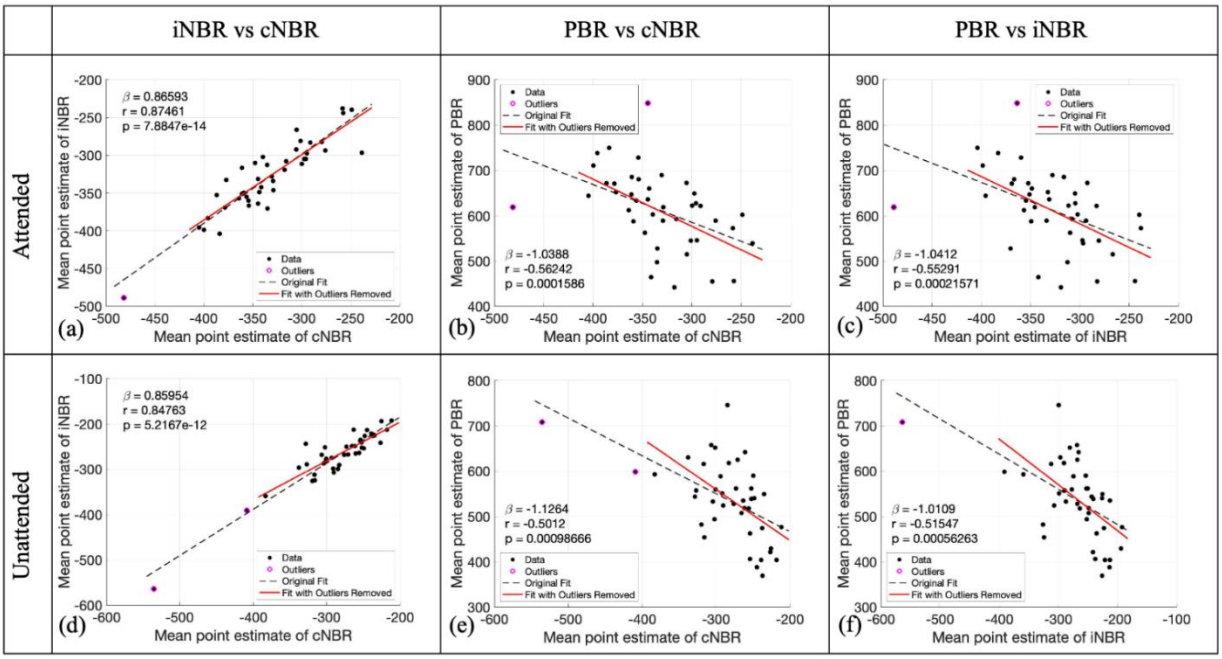

Figure 6.6: Comparisons between the mean strengths of the BOLD responses to the attended (a, b, and c) and unattended stimuli (d, e, and f). Regression lines including all data points are depicted by dashed black lines, while the solid red lines depict the regression lines in the absence of the few outlier data points. As quantitatively shown here, independent of the attention case, the correlation between iNBR and cNBR is significantly higher than those with PBR.

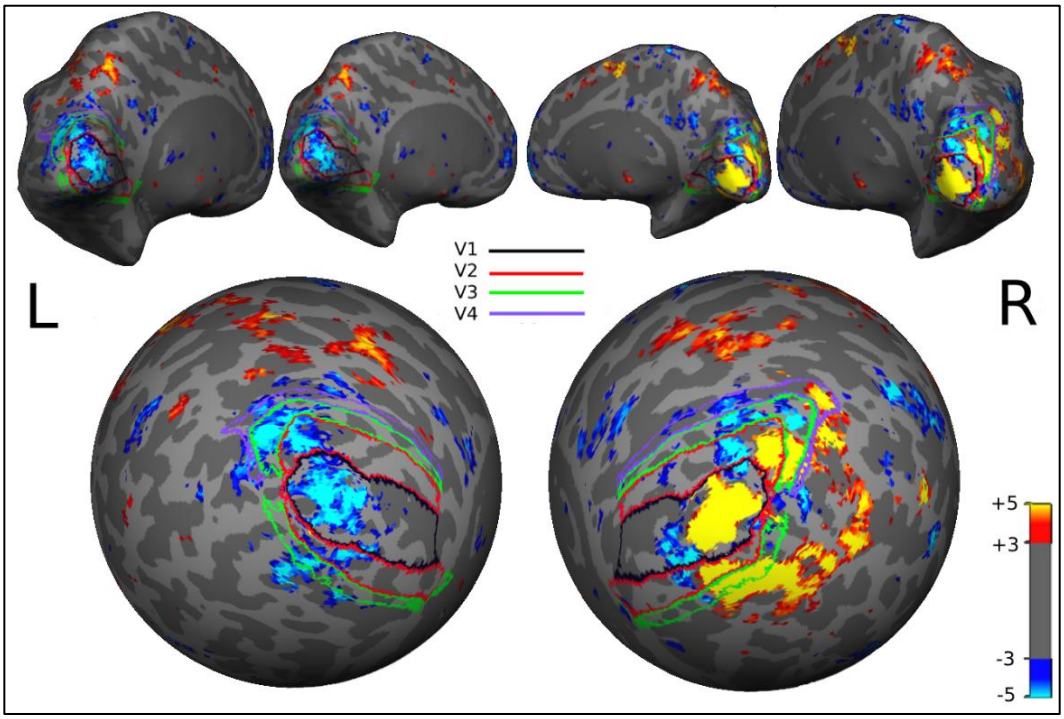

Figure 6.7: Visualization of spatial similarities between the BOLD response patterns. The margins for different visual areas are shown in different colors.

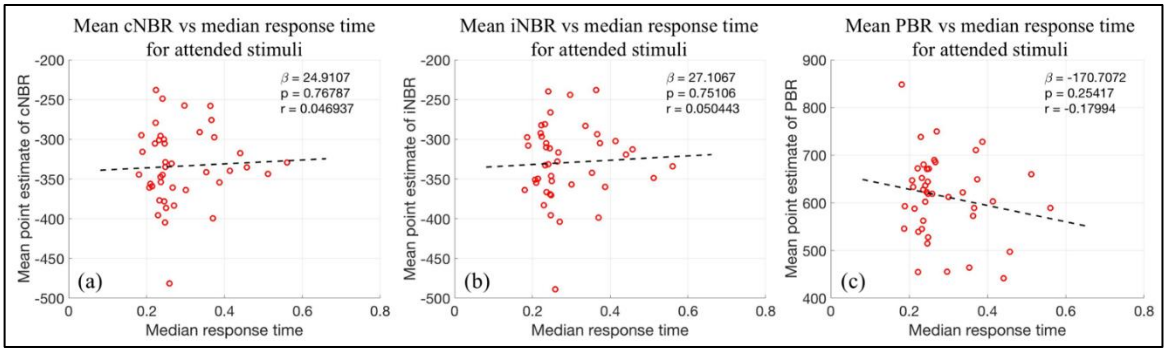

Figure 6.8: Relationship of the amplitudes of the a) cNBR, b) iNBR, and c) PBR with task performance. The dashed lines show the trend in the data.

6.3.6 Attention Modulation of PBR, cNBR, and iNBR

In this analysis, we investigated the effect of attention on the BOLD responses. Our null hypothesis was that there would be no difference between extracted HRFs (resulting from visual stimulation) between the attended and unattended cases at each time point. Results are illustrated in Fig. 6.9. As shown here, attention did not significantly modulate PBR, except at t=2 seconds post-stimulation where a significant change between the amplitudes of attended and unattended HRFs was observed. Meanwhile, the amplitude of both NBRs' HRFs significantly increased around 3-4 seconds after the stimulus began as a result of attention (cNBR: 18.89%, $t = 1.4977$, $p > 0.1380$; iNBR: 23.28%, $t = 3.3751$, $p < 0.0011$). Similarly, attention modulated the return to baseline times for both NBR HRFs by reducing them (cNBR: Δt=-0.6921 s, $t = -1.3098$, $p > 0.1939$; iNBR: Δt= -1.7913 s, $t = -2.6791$, $p < 0.0090$). While there are some key differences between the iNBR and cNBR HRFs' modulations by attention in terms of reaching the statistical significance level, overall, they seem to share a similar trend of attention dependency (Fig. 6.9 b and c, highlighted by the red shades). This modulation trend seems to be different than the effect of attention on PBR's HRF (Fig. 6.9 a). These results suggest that the potential mechanism underlying iNBR and cNBR is likely to be different from that of PBR.

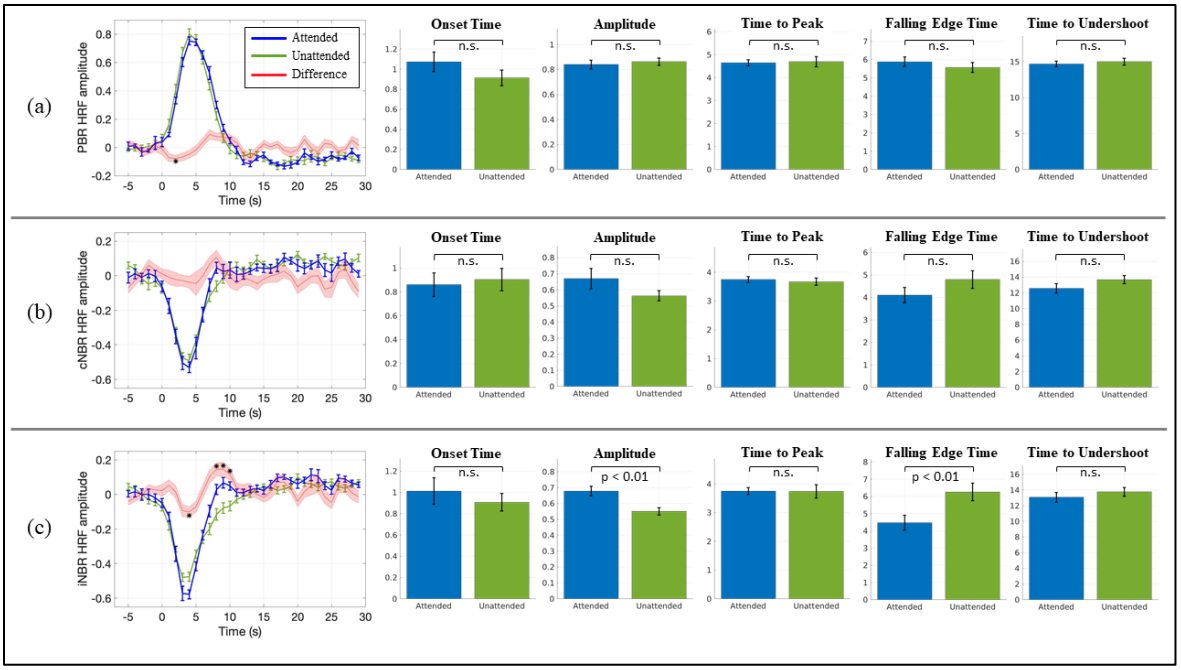

Figure 6.9: The effect of attention on a) PBR, b) cNBR, and c) iNBR. PBR gets affected by attention only at t=2 s while cNBR and iNBR show similar modulation trends by attention. Statistically significant differences are marked with asterisk symbols (* $\alpha = 0.05$, Bonferroni correction).

6.4. Discussion and Conclusion

In this chapter, we studied the characteristics of two NBRs induced simultaneously by unilateral visual stimulation in the human visual cortex contralaterally and ipsilaterally. To our knowledge, no prior study has detected and characterized the cNBR and iNBR evoked simultaneously following the same unilateral visual stimulation. First, we showed that the strengths of both NBRs increased linearly with visual stimulus duration. Although linearity of individual NBR types has been investigated before (Shmuel et al. 2002; Klingner et al. 2010), however, the linearity of concurrent occurrence of iNBR and cNBR due to unilateral stimulation is under investigated. The linearity assessment is of extreme importance since most fMRI analyses use the linearity assumption to extract the BOLD signals (e.g., general linear modeling).

Furthermore, the analysis of the HRFs extracted from the cNBR and iNBR revealed no significant difference in amplitudes and dynamics after the multiple comparison correction.

However, this was not the case for PBR's HRF which showed significant differences with the HRFs of the two NBRs. Similar findings have been reported in the literature on the difference between PBR and NBR HRFs. For example, (Shmuel et al. 2002) reported a significant difference between the falling edge times of PBR and cNBR HRFs extracted from the visual cortex of the human brain. Moreover, our results were consistent with (Shmuel et al. 2002) in that there was a significant correlation between the expression of cNBR and PBR. Interestingly, in our analysis, the correlation between cNBR and iNBR was significantly higher than the correlation between the BOLD responses of the same brain hemisphere (*i.e.*, PBR and cNBR). This novel observation suggests that the two NBRs, despite being separated by two brain hemispheres, are likely to be generated by a common process. Furthermore, we illustrated that the spatial similarities between the fMRI patterns of iNBR and both cNBR (Dice score below 5%) and PBR (Dice score below 6%) are very low. This evidence illustrates that the callosal pathways, could not be responsible for simultaneous observation of tightly coupled cNBR and iNBR, otherwise there would have been a higher spatial similarity score between the two NBRs than between the iNBR and PBR.

Furthermore, we did not find a significant correlation between task performance and the two NBRs. Although behavioral correlates of PBR have been reported (Pessoa et al. 2002), there is no consensus on the behavioral correlates of NBRs. Some studies suggest a significant correlation between task performance and the amplitude of iNBR (Kastrup et al. 2008). In contrast, under similar experimental conditions as in (Kastrup et al. 2008), authors in (Schäfer et al. 2012) did not find such a statistically significant relationship. Further experiments with larger sample sizes are needed to thoroughly investigate the behavioral correlates of the two NBRs.

We also found that attending to visual stimuli led to a very small but significant increase in the ipsilateral NBR. Similarly, other studies have reported an increase in the overall NBR size

in response to increased attention (Bressler et al. 2013; Heinemann, Kleinschmidt, and Müller 2009; Müller and Kleinschmidt 2004). In addition to the magnitude, by analyzing the dynamics of the extracted HRFs time-series corresponding to the two NBRs, we observed similar patterns of differences between the two in the presence and absence of attention. Similar findings have been reported about the attention modulation on NBR HRFs in macaque monkeys' visual cortex (Shmuel et al. 2006).

Considering the similarities in the properties and characteristics of the two NBRs investigated in this research, we speculate that they are the result of the same process. In other words, our findings suggest that the cNBR and iNBR are associated with similar neural and/or vascular mechanisms. Since the characteristics of the two NBRs were mostly different from PBR, this mechanism is perhaps different from the PBR's underlying mechanism, and it may involve subcortical brain structures with projections into both hemispheres.

Conclusion

Summary

In this dissertation, we applied Artificial Intelligence and Signal Processing techniques to large cohorts of MRI and CT images with the goal of intelligent interpretation of complex visual patterns in imaging data. As more MRI and CT scans are conducted every day, it is essential to design and develop frameworks that enable automatic, reliable, and fast processing of such images for clinical and research purposes. Doing so results in substantial savings in costs and experts' time. Additionally, Artificial Intelligence can detect patterns in the imaging data that might not necessarily be visible to the human eye. This results in less biased and less subjective interpretation of the data.

In particular, in this dissertation we applied Artificial Intelligence and Signal Processing methods to four medical imaging problems, namely: 1) Weakly-supervised automated quality assessment of dMRI volumes; 2) Weakly-supervised automated major artifact type identification in dMRI poor-quality volumes; 3) Supervised automated labeling of spatially-informed lung texture patterns (sLTPs) associated with Emphysema subtypes on lung CT images; and 4) Detection and characterization of ipsilateral and contralateral negative BOLD responses in human visual cortex on fMRI scans.

In the first problem, through the use of weakly-supervised learning strategies, we demonstrated that dMRI volumes can be reliably, quickly, and automatically quality controlled into good and poor-quality volumes. Specifically, we developed a 2-step framework for this task. In the first step, we designed and implemented two convolutional neural networks (CNNs) to predict the quality labels associated with ROIs of two different sizes (i.e., small and large) extracted from the dMRI volume under investigation. In the second step, we utilized these

predicted labels in a tailored voting algorithm to predict the final quality label associated with the whole dMRI volume under investigation. Our results demonstrated the practicality of our proposed artificial intelligence-based framework for fast and automatic quality control of dMRI volumes.

We motivated the second problem by acknowledging that a binary quality assessment is a necessary but not sufficient step prior to carrying out any further analysis on medical images. Since acquiring medical images such as dMRI are expensive, it is often beneficial to detect the type of the existing artifact in poor-quality volumes, and try to restore the image (if possible). Hence, in order to address this problem, by using weakly-supervised learning strategies, we designed and implemented a 2-step framework for automatic and reliable artifact type classification in dMRI volumes. In the first step, we developed a Squeeze-and-Excitation (SE) CNN to predict quality labels of N mutually exclusive and collectively exhaustive (MECE) slabs extracted from the dMRI whole volume under examination. In the second step, we used the predicted slab-level artifact labels to predict the volume-level artifact class using a voting algorithm. We tested our model on two different cohorts of dMRI data and demonstrated the practicality of our framework for automatic four class artifact classification.

In the third problem, we transited from MRI modality to CT. Specifically, we focused on large cohorts of lung CT scans of human subjects with history of Emphysema disease. Through the use of supervised learning approach, we designed a framework for automatic, and cohort-invariant labeling of ten class of sLTPs associated with six novel CT Emphysema subtypes (CTES). In particular, we designed and implemented a SE-CNN to predict the sLTP class labels of small ROIs randomly sampled from lung CT images. Following this step, by using the mapping between sLTPs and CTES, we used the predicted sLTP labels of sampled ROIs to perform Emphysema subtyping and predict their CTES class label. We trained, validated, and tested our

model on one cohort, namely SPIROMICS V1, and then in order to show robustness with respect to change in cohort, we tested our model (as is, without any modifications or parameter tuning) on MESA-Lung exam 5. Our results demonstrated robustness of our proposed sLTP and CTES classifier with respect to different cohorts and scanner types.

Finally, in the fourth and last problem, we focused on two types of negative BOLD responses (NBR) elicited in human brain during unilateral visual stimulation that are captured via fMRI scans. In particular, we designed an event-related fMRI experiment with two unilateral sensory stimulations (i.e., visual and audio stimuli) that resulted in sustained ipsilateral negative BOLD response (iNBR) and contralateral negative BOLD response (cNBR) in human brain. We focused on iNBR and cNBR due to visual stimulation and conducted several experiments to visualize and characterize these two responses. In particular, through the use of mathematical modeling and signal processing techniques, we showed the linearity of both NBRs with respect to stimulus duration. We then extracted the hemodynamic response functions (HRFs) of the two NBRs and demonstrated that the two HRFs have no statistically significant difference in their overall shape and dynamic. Next, we showed that the two NBRs are tightly coupled as quantified by their strong and statistically significant Pearson correlation coefficients. Following that, we tested the relationship between the fMRI task performance and the two NBRs and realized that there is no association between the two. In order to investigate if bilateral structural connections such as Callosal pathways could account for the above-mentioned similarities between the two NBRs, we tested the spatial similarity between the response patterns of the two NBRs. The result of this analysis provided evidence that Callosal pathways are most likely not the underlying process behind iNBR and cNBR. Finally, we investigated the effect of attention on the HRFs of iNBR and cNBR and realized that attention modulates the two HRFs a few second following the

onset of stimulation by increasing their magnitudes. Putting all results together, we concluded that there is a common mechanism underlying iNBR and cNBR and this mechanism is potentially different from the one giving rise to the positive BOLD response.

Future Directions

Despite the contributions made in this dissertation, there still remains room for more contribution to further advance the field. In this section, we briefly discuss some potential future directions for each one of the four projects discussed in this dissertation.

Future directions for Problem 1 and 2: One avenue for future directions for these two problems, is to learn artifact finger prints directly from the K-space (i.e., raw MRI measurements space containing spatial frequencies) instead of the image domain. K-space contains all information needed to construct an MR image. Hence, in theory, one should be able to detect artifacts directly from K-space prior to image reconstruction. Since K-space contains complex numbers, one approach would be to either design CNNs that directly operate with complex-valued numbers, or to treat the real and imaginary parts of the measurements as two separate channels that are related to each other by satisfying the magnitude constraint (i.e., $\|C\| = \sqrt{(Re\{C\})^2 + (Im\{C\})^2}$ where C is a complex number, Re{} denotes the real part, and Im{} denotes the imaginary part of C). Detecting artifacts directly from K-space prior to image reconstruction is beneficial in the sense that one does not need to wait for the image to be fully reconstrued to perform quality control. This allows for integration of an intelligent real-time quality control tool with the MRI scanner. Hence, in the event that the acquired MR scan is contaminated with artifact(s), the MR technician is able to conduct another scan while the patient (or subject in research settings) is still in the scanner.

Another path for future work is to expand the number of artifact classes for the artifact classifier that was designed and implemented in this thesis. In particular, the fourth class (i.e., MRI miscellaneous artifacts) served as an umbrella artifact category. One can further divide this class to a number of sub-classes corresponding to different artifacts such as Eddy currents, susceptibility artifacts, etc.

Future directions for Problem 3: In this project we developed a classifier for sLTP and CTES labeling using only the information available in a lung CT scan. However, lung CT scans are not the only source of information for investigating COPD diseases such as Emphysema. The genomic data contain a great deal of information on risk factors associated with COPD diseases. One future direction would be to use the genomic information in combination with the CT imaging data for studying Emphysema subtyping and progression. One way to approach this would be to design a CNN that takes both CT images and genomic data as inputs to process the emphysema related information. Another future direction is to investigate the longitudinal progression of sLTPs and CTES and to see whether an artificial intelligence-based model is capable of predicting the transition between different sLTPs as Emphysema progresses over time in the lung.

Future directions for Problem 4: In this project, we investigated the properties of iNBR and cNBR and concluded that the two possibly have the same underlying mechanism. However, we were unable to identify what exactly this mechanism could be. Identifying such a mechanism requires more experiments perhaps by including data from other modalities (such as EEG recordings). Since fMRI is an indirect way of measuring neuronal activities, it would be better to combine fMRI data with other direct measurements of neuronal activity to investigate the underlying mechanism of the two NBRs. Doing so can tell us more about the possibility of having a pure or partial neuronal mechanism underlying iNBR and cNBR.

Bibliography

Abadi, Martín, et al. (2016). "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint* arXiv:1603.04467 (2016).

Ahmad, Adnan, et al. (2021). "3D-QCNet--A Pipeline for Automated Artifact Detection in Diffusion MRI images." *arXiv preprint* arXiv:2103.05285 (2021).

Akkus, Zeynettin, et al. (2017). "Deep learning for brain MRI segmentation: state of the art and future directions." *Journal of digital imaging* 30.4 (2017): 449-459.

Alexander, Andrew L., et al. (2007). "Diffusion tensor imaging of the brain." *Neurotherapeutics* 4.3 (2007): 316-329.

Alexander, Lindsay M., et al. (2017). "An open resource for transdiagnostic research in pediatric mental health and learning disorders." *Scientific data* 4.1 (2017): 1-26

Alfaro-Almagro, Fidel, et al. (2018). "Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank." *Neuroimage* 166 (2018): 400-424.

Anaya-Isaza, Andrés, Leonel Mera-Jiménez, and Martha Zequera-Diaz (2021). "An overview of deep learning in medical imaging." *Informatics in Medicine* Unlocked 26 (2021): 100723.

Anderson Jr, Augustus E., et al. (1964). "Emphysema in lung macrosections correlated with smoking habits." *Science* 144.3621 (1964): 1025-1026.

Andersson, Jesper L R et al. (2016). "Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images." *NeuroImage* vol. 141 (2016): 556-572. doi:10.1016/j.neuroimage.2016.06.058

Andersson, Jesper LR, Stefan Skare, and John Ashburner (2003). "How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging." *Neuroimage* 20.2 (2003): 870-888.

Asherov, Marina, Idit Diamant, and Hayit Greenspan (2014). "Lung texture classification using bag of visual words." *Medical Imaging 2014: Computer-Aided Diagnosis*. Vol. 9035. SPIE, 2014.

Baliyan, V., et al. (2016). “Diffusion weighted imaging: technique and applications”. *World journal of radiology*, 2016. 8(9): p. 785.

Bammer, Roland, et al. (2003). “Analysis and generalized correction of the effect of spatial gradient field distortions in diffusion-weighted imaging.” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 50.3 (2003): 560-569.

Basser, Peter J., and Derek K. Jones (2002). “Diffusion-tensor MRI: theory, experimental design and data analysis–a technical review.” *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 15.7-8 (2002): 456-467.

Bastiani, Matteo, et al. (2019). “Automated quality control for within and between studies diffusion MRI data using a non-parametric framework for movement and distortion correction.” *Neuroimage* 184 (2019): 801-812

Bercovich, Eyal, and Marcia C. Javitt (2018). “Medical imaging: from roentgen to the digital revolution, and beyond.” *Rambam Maimonides medical journal* 9.4 (2018).

Bermejo-Peláez, David, Raúl San José Estepar, and María J. Ledesma-Carbayo (2018). “Emphysema classification using a multi-view convolutional network.” *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018.

Bild, Diane E., et al. (2002). “Multi-ethnic study of atherosclerosis: objectives and design.” *American journal of epidemiology* 156.9 (2002): 871-881.

Binder, Polina, et al. (2016). “Unsupervised discovery of emphysema subtypes in a large clinical cohort.” *International Workshop on Machine Learning in Medical Imaging*. Springer, Cham, 2016.

Boas, David A., et al. (2008). “A vascular anatomical network model of the spatio-temporal response to brain activation.” *Neuroimage* 40.3 (2008): 1116-1129.

Bocci, Tommaso, et al. (2014). “Visual callosal connections: role in visual processing in health and disease.” *Reviews in the Neurosciences* 25.1 (2014): 113-127.

Boynton, Geoffrey M., et al. (1996). “Linear systems analysis of functional magnetic resonance imaging in human V1.” *Journal of Neuroscience* 16.13 (1996): 4207-4221.

Bressler, David W., et al. (2013). “Visual spatial attention enhances the amplitude of positive and negative fMRI responses to visual stimulation in an eccentricity-dependent manner.” *Vision research* 85 (2013): 104-112.

Buzug, Thorsten M. (2011). “Computed tomography.” *Springer Handbook of Medical Technology*. Springer, Berlin, Heidelberg, 2011. 311-342.

Casey, B. J., et al. (2018). “The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites.” *Developmental cognitive neuroscience* 32 (2018): 43-54.

Chollet, F., et al. *Keras*. (2015). https://github.com/fchollet/keras.

Couper, David, et al. (2014). “Design of the subpopulations and intermediate outcomes in COPD study (SPIROMICS).” *Thorax* 69.5 (2014): 492-495.

Dahl, Morten, et al. (2002). “Change in lung function and morbidity from chronic obstructive pulmonary disease in α1-antitrypsin MZ heterozygotes: a longitudinal study of the general population.” *Annals of internal medicine* 136.4 (2002): 270-279.

Depeursinge, Adrien, et al. (2014). “Three-dimensional solid texture analysis in biomedical imaging: review and opportunities.” *Medical image analysis* 18.1 (2014): 176-196.

Druzhkov, P. N., and V. D. Kustikova (2016). “A survey of deep learning methods and software tools for image classification and object detection.” *Pattern Recognition and Image Analysis* 26.1 (2016): 9-15.

Esteva, Andre, et al. (2017). “Dermatologist-level classification of skin cancer with deep neural networks.” *nature* 542.7639 (2017): 115-118.

Ettehadi, Nabil, et al. (2021). “Automatic Volumetric Quality Assessment of Diffusion MR Images via Convolutional Neural Network Classifiers.” *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021.

Ettehadi, Nabil, et al. (2022). “Automated Multiclass Artifact Detection in Diffusion MRI Volumes via 3D Residual Squeeze-and-Excitation Convolutional Neural Networks.” *Frontiers in Human Neuroscience* (2022).

Evans, Alan C., et al. (1993). “3D statistical neuroanatomical models from 305 MRI volumes.” *1993 IEEE conference record nuclear science symposium and medical imaging conference*. IEEE, 1993.

Fabri, Mara, et al. (2011). “Topographical organization of human corpus callosum: an fMRI mapping study.” *Brain research* 1370 (2011): 99-111.

Fantini, Irene, et al. (2018). “Automatic detection of motion artifacts on MRI using Deep CNN.” 2018 *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, 2018.

Fischl, Bruce, et al. (2002). “Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain.” *Neuron* 33.3 (2002): 341-355.

Fischl, Bruce, et al. (2004). “Automatically parcellating the human cerebral cortex.” *Cerebral cortex* 14.1 (2004): 11-22.

Friston, Karl J., et al. (1998). “Nonlinear event-related responses in fMRI.” *Magnetic resonance in medicine* 39.1 (1998): 41-52.

Gangeh, Mehrdad J., et al. (2010). “A texton-based approach for the classification of lung parenchyma in CT images.” *International conference on medical image computing and computer-assisted intervention*. Springer, Berlin, Heidelberg, 2010.

Ginsburg, Shoshana B., et al. (2012). “Automated texture-based quantification of centrilobular nodularity and centrilobular emphysema in chest CT images.” *Academic Radiology* 19.10 (2012): 1241-1251.

Glorot, Xavier, and Yoshua Bengio (2010). “Understanding the difficulty of training deep feedforward neural networks.” *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010.

Goense, Jozien, Hellmut Merkle, and Nikos K. Logothetis (2012). “High-resolution fMRI reveals laminar differences in neurovascular coupling between positive and negative BOLD responses.” *Neuron* 76.3 (2012): 629-639.

Good, I. J. (1952). “Rational Decisions.” *Journal of the Royal Statistical Society*. Series B (Methodological), vol. 14, no. 1, 1952, pp. 107–114. JSTOR, www.jstor.org/stable/2984087

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press, 2016.

Goutte, Cyril, Finn Aarup Nielsen, and K. H. Hansen (2000). “Modeling the hemodynamic response in fMRI using smooth FIR filters.” *IEEE transactions on medical imaging* 19.12 (2000): 1188-1201.

Graham, Mark S., Ivana Drobnjak, and Hui Zhang (2018). “A supervised learning approach for diffusion MRI quality control with minimal training data.” *NeuroImage* 178 (2018): 668-676.

Gulshan, Varun, et al. (2016). “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.” *Jama* 316.22 (2016): 2402-2410.

Haddad, Seyyed MH, et al. (2019). “Comparison of quality control methods for automated diffusion tensor imaging analysis pipelines.” *PloS one* 14.12 (2019): e0226715.

Häme, Yrjö, et al. (2014). “Adaptive quantification and longitudinal analysis of pulmonary emphysema with a hidden Markov measure field model.” *IEEE transactions on medical imaging* 33.7 (2014): 1527-1540.

Häme, Yrjö, et al. (2015). “Sparse sampling and unsupervised learning of lung texture patterns in pulmonary emphysema: MESA COPD study.” *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2015.

Han, Jun, and Claudio Moraga (1995). "The influence of the sigmoid function parameters on the speed of backpropagation learning." *International workshop on artificial neural networks.* Springer, Berlin, Heidelberg, 1995.

Han, Su-Hyun, et al. (2018). "Artificial neural network: understanding the basic concepts without mathematics." *Dementia and Neurocognitive Disorders* 17.3 (2018): 83-89.

Harel, Noam, et al. (2002). "Origin of negative blood oxygenation level—dependent fMRI signals." *Journal of cerebral blood flow & metabolism* 22.8 (2002): 908-917.

He, Kaiming, et al. (2016). "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016

Heinemann, Linda, Andreas Kleinschmidt, and Notger G. Müller (2009). "Exploring BOLD changes during spatial attention in non-stimulated visual cortex." *PLoS One* 4.5 (2009): e5560.

Herath, Samitha, Mehrtash Harandi, and Fatih Porikli (2017). "Going deeper into action recognition: A survey." *Image and vision computing* 60 (2017): 4-21.

Hoy, Andrew R., et al. (2017). "Microstructural white matter alterations in preclinical Alzheimer's disease detected using free water elimination diffusion tensor imaging." *PloS one* 12.3 (2017): e0173982.

Hu, Dewen, and Liangming Huang (2015). "Negative hemodynamic response in the cortex: evidence opposing neuronal deactivation revealed via optical imaging and electrophysiological recording." *Journal of Neurophysiology* 114.4 (2015): 2152-2161.

Hu, Jie, Li Shen, and Gang Sun (2018). "Squeeze-and-excitation networks." *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141. 2018.

Huisman, T. A. G. M. (2010). "Diffusion-weighted and diffusion tensor imaging of the brain, made easy." *Cancer Imaging* 10.1A (2010): S163.

Huisman, Thierry AGM. (2003). "Diffusion-weighted imaging: basic concepts and application in cerebral stroke and head trauma." *European radiology* 13.10 (2003): 2283-2297.

Hüppi, Petra S., and Jessica Dubois (2006). "Diffusion tensor imaging of brain development." *Seminars in Fetal and Neonatal Medicine*. Vol. 11. No. 6. WB Saunders, 2006.

Iglesias, Juan Eugenio, et al. (2017). "Retrospective head motion estimation in structural brain MRI with 3D CNNs." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2017.

J. Ross et al. (2017). "A bayesian nonparametric model for disease subtyping: Application to emphysema phenotypes," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 343–354, 2017.

Jenkinson, Mark, et al. (2012). "Fsl." *Neuroimage* 62.2 (2012): 782-790.

Jiang, Hangyi, et al. (2006). “DtiStudio: resource program for diffusion tensor computation and fiber bundle tracking.” *Computer methods and programs in biomedicine* 81.2 (2006): 106-116.

K. Aoshiba, N. Yokohori, and A. Nagai (2003). “Alveolar wall apoptosis causes lung destruction and emphysematous changes,” *Am. J. Respir. Cell Mol. Biol*., vol. 28, no. 5, pp. 555–562, 2003.

Kastrup, Andreas, et al. (2008). “Behavioral correlates of negative BOLD signal changes in the primary somatosensory cortex.” *Neuroimage* 41.4 (2008): 1364-1371.

Kelly, Christopher, et al. (2017). “Transfer learning and convolutional neural net fusion for motion artefact detection.” *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine*, Honolulu, Hawaii. Vol. 3523. 2017.

Kingma, Diederik P., and Jimmy Ba (2014). “Adam: A method for stochastic optimization.” *arXiv preprint* arXiv:1412.6980 (2014).

Kingma, Diederik P., and Jimmy Ba. (2014). “Adam: A method for stochastic optimization.” *arXiv preprint* arXiv:1412.6980 (2014).

Klingner, Carsten M., et al. (2010). “Dependence of the negative BOLD response on somatosensory stimulus intensity.” *Neuroimage* 53.1 (2010): 189-195.

Krenker, Andrej, Janez Bešter, and Andrej Kos (2011). “Introduction to the artificial neural networks.” *Artificial Neural Networks: Methodological Advances and Biomedical Applications*. InTech (2011): 1-18.

Krogsrud, Stine K., et al. (2016). “Changes in white matter microstructure in the developing brain—A longitudinal diffusion tensor imaging study of children from 4 to 11 years of age.” *Neuroimage* 124 (2016): 473-486.

Krupa, Katarzyna, and Monika Bekiesińska-Figatowska (2015). “Artifacts in magnetic resonance imaging.” *Polish journal of radiology* 80 (2015): 93.

L. Sørensen, S. B. Shaker, and M. De Bruijne (2010). “Quantitative analysis of pulmonary emphysema using local binary patterns,” *IEEE Trans. Med. Imaging*, vol. 29, no. 2, pp. 559–569, 2010.

Ladouceur, Cecile D., et al. (2012). “White matter development in adolescence: the influence of puberty and implications for affective disorders.” *Developmental cognitive neuroscience* 2.1 (2012): 36-54.

Lakhani, Paras, et al. (2018). “Machine learning in radiology: applications beyond image interpretation.” *Journal of the American College of Radiology* 15.2 (2018): 350-359.

Le Bihan, Denis, et al. (1986). “MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders.” *Radiology* 161.2 (1986): 401-407.

Le Bihan, Denis, et al. (2006). “Artifacts and pitfalls in diffusion MRI.” *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 24.3 (2006): 478-488.

Li, Haipeng, and Ramakrishnan Mukundan (2020). “Robust texture features for emphysema classification in CT images.” *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

Liu, Bilan, Tong Zhu, and Jianhui Zhong (2015). “Comparison of quality control software tools for diffusion tensor imaging.” *Magnetic resonance imaging* 33.3 (2015): 276-285.

Liu, Jingyu, et al. (2019). “Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision.” *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

Liu, Xiaoqing, et al. (2021). “Advances in deep learning-based medical image analysis.” *Health Data Science* 2021 (2021).

Liu, Yadong, et al. (2011). “Sustained negative BOLD response in human fMRI finger tapping task.” *PloS one* 6.8 (2011): e23839.

Lo Buono, Viviana, et al. (2020). “Diffusion tensor imaging of white matter degeneration in early stage of Alzheimer’s disease: a review.” *International Journal of Neuroscience* 130.3 (2020): 243-250.

Logothetis, N.K. (2008). “What we can do and what we cannot do with fMRI”. *Nature*, 453(7197), p.869. (2008).

Maas, Benjamin, Erfan Zabeh, and Soroush Arabshahi (2021). “QuickTumorNet: Fast Automatic Multi-Class Segmentation of Brain Tumors.” *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2021.

Matheny, Michael, et al. (2019). “Artificial intelligence in health care: the hope, the hype, the promise, the peril.” *NAM Special Publication*. Washington, DC: National Academy of Medicine (2019): 154.

McRobbie, Donald W., et al. (2017). “MRI from Picture to Proton”. *Cambridge university press*, 2017.

Mets, O. M., et al. (2012). “Quantitative computed tomography in COPD: possibilities and limitations.” *Lung* 190.2 (2012): 133-145.

Minaee, Shervin, et al. (2021). “Image segmentation using deep learning: A survey.” *IEEE transactions on pattern analysis and machine intelligence* (2021).

Müller, Notger G., and Andreas Kleinschmidt (2004). “The attentional ‘spotlight’s’ penumbra: center-surround modulation in striate cortex.” *Neuroreport* 15.6 (2004): 977-980.

Mullinger, Karen J., et al. (2014). “Evidence that the negative BOLD response is neuronal in origin: a simultaneous EEG–BOLD–CBF study in humans.” *Neuroimage* 94 (2014): 263-274.

Nair, Vinod, and Geoffrey E. Hinton (2010). “Rectified linear units improve restricted boltzmann machines.” *Icml*. 2010.

Ogawa, Seiji, et al. (1992). “Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging.” *Proceedings of the National Academy of Sciences* 89.13 (1992): 5951-5955.

Oguz, Ipek, et al. (2014). “DTIPrep: quality control of diffusion-weighted images.” *Frontiers in neuroinformatics* 8 (2014): 4.

Papanicolas, Irene, Liana R. Woskie, and Ashish K. Jha (2018). “Health care spending in the United States and other high-income countries.” *Jama* 319.10 (2018): 1024-1039.

Parker, David B., and Qolamreza R. Razlighi (2019). “Task-evoked negative BOLD response and functional connectivity in the default mode network are representative of two overlapping but separate neurophysiological processes.” *Scientific reports* 9.1 (2019): 1-17.

Parker, David, Xueqing Liu, and Qolamreza R. Razlighi (2017). “Optimal slice timing correction and its interaction with fMRI parameters and artifacts.” *Medical image analysis* 35 (2017): 434-445.

Peng, Liying, et al. (2018). “Classification of pulmonary emphysema in CT images based on multi-scale deep convolutional neural networks.” *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018.

Peng, Liying, et al. (2019a). “Classification and quantification of emphysema using a multi-scale residual network.” *IEEE journal of biomedical and health informatics* 23.6 (2019): 2526-2536.

Peng, Liying, et al. (2019b). “Semi-supervised learning for semantic segmentation of emphysema with partial annotations.” *IEEE Journal of Biomedical and Health Informatics* 24.8 (2019): 2327-2336.

Pesapane, Filippo, Marina Codari, and Francesco Sardanelli (2018). “Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine.” *European radiology experimental* 2.1 (2018): 1-10.

Pessoa, Luiz, et al. (2002). “Neural correlates of visual working memory: fMRI amplitude predicts task performance.” *Neuron* 35.5 (2002): 975-987.

Pierpaoli, C., et al. (2010). “TORTOISE: an integrated software package for processing of diffusion MRI data.” *ISMRM 18th annual meeting*. Vol. 1597. 2010.

Reuter, Martin, et al. (2015). “Head motion during MRI acquisition reduces gray matter volume and thickness estimates.” *Neuroimage* 107 (2015): 107-115.

Roalf, David R et al. (2015). “White matter microstructure in schizophrenia: associations to neurocognition and clinical symptomatology.” *Schizophrenia research* vol. 161,1 (2015): 42-9. doi:10.1016/j.schres.2014.09.026

Roalf, David R., et al. (2016). “The impact of quality assurance assessment on diffusion tensor imaging outcomes in a large-scale population-based cohort.” *Neuroimage* 125 (2016): 903-919.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional networks for biomedical image segmentation.” *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

Roy, Abhijit Guha, et al. (2019). “QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy.” *NeuroImage* 186 (2019): 713-727.

Samani, Zahra Riahi, et al. (2020). “QC-Automator: Deep learning-based automated quality control for diffusion mr images.” *Frontiers in neuroscience* 13 (2020): 1456.

Schäfer, Katharina, et al. (2012). “Negative BOLD signal changes in ipsilateral primary somatosensory cortex are associated with perfusion decreases and behavioral evidence for functional inhibition.” *Neuroimage* 59.4 (2012): 3119-3127.

Shmuel, A., et al. (2003). “Negative BOLD response ipsi-lateral to the visual stimulus: Origin is not blood stealing.” *9th International Conference on Functional Mapping of the Human Brain (HBM 2003)*. Academic Press, 2003.

Shmuel, Amir, et al. (2002). “Sustained negative BOLD, blood flow and oxygen consumption response and its coupling to the positive response in the human brain.” *Neuron* 36.6 (2002): 1195-1210.

Shmuel, Amir, et al. (2006). “Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1.” *Nature neuroscience* 9.4 (2006): 569-577.

Sieren, Jered P., et al. (2016). “SPIROMICS protocol for multicenter quantitative computed tomography to phenotype the lungs.” *American journal of respiratory and critical care medicine* 194.7 (2016): 794-806

Simmonds, Daniel J., et al. (2014). “Developmental stages and sex differences of white matter and behavioral development through adolescence: a longitudinal diffusion tensor imaging (DTI) study.” *Neuroimage* 92 (2014): 356-368.

Skourt, Brahim Ait, Abdelhamid El Hassani, and Aicha Majda (2018). “Lung CT image segmentation using deep neural networks.” *Procedia Computer Science* 127 (2018): 109-113.

Smith, A. T., K. D. Singh, and M. W. Greenlee (2000). “Attentional suppression of activity in the human visual cortex.” *Neuroreport* 11.2 (2000): 271-278.

Smith, Andrew T., Adrian L. Williams, and Krishna D. Singh (2004). “Negative BOLD in the visual cortex: evidence against blood stealing.” *Human brain mapping* 21.4 (2004): 213-220.

Smith, S.M., et al. (2004). “Advances in functional and structural MR image analysis and implementation as FSL.” *Neuroimage* 23 (2004): S208-S219.

Spanhol, Fabio A., et al. (2017). “Deep features for breast cancer histopathological image classification.” *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017.

Stefanovic, Bojana, Jan M. Warnking, and G. Bruce Pike (2004). “Hemodynamic and metabolic responses to neuronal inhibition.” *Neuroimage* 22.2 (2004): 771-778.

Stejskal, Edward O., and John E. Tanner (1965). “Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient.” *The journal of chemical physics* 42.1 (1965): 288-292.

Sutskever, Ilya, et al. (2013). “On the importance of initialization and momentum in deep learning.” *International conference on machine learning*. PMLR, 2013.

Ting, Daniel Shu Wei, et al. (2017). “Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes.” *Jama* 318.22 (2017): 2211-2223.

Tønnesen, Siren, et al. (2018). "White matter aberrations and age-related trajectories in patients with schizophrenia and bipolar disorder revealed by diffusion tensor imaging." *Scientific reports* 8.1 (2018): 1-14.

Tournier, J-Donald, et al. (2019). “MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation.” *NeuroImage* 202 (2019): 116137.

Van Dijk, Koene RA, Mert R. Sabuncu, and Randy L. Buckner (2012). “The influence of head motion on intrinsic functional connectivity MRI.” *Neuroimage* 59.1 (2012): 431-438.

Van Essen, D C et al. (2012). “The Human Connectome Project: a data acquisition perspective.” *NeuroImage* vol. 62,4 (2012): 2222-31. doi:10.1016/j.neuroimage.2012.02.018

Wang, Chu-ran, et al. (2020). “BR-GAN: bilateral residual generating adversarial network for mammogram classification.” *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2020.

Wang, Mei, and Weihong Deng (2021). “Deep face recognition: A survey.” *Neurocomputing* 429 (2021): 215-244.

Wilson, Edwin B. (1927). “Probable inference, the law of succession, and statistical inference.” *Journal of the American Statistical Association* 22.158 (1927): 209-212.

Woolrich, Mark W., Timothy EJ Behrens, and Stephen M. Smith (2004). “Constrained linear basis sets for HRF modelling using Variational Bayes.” *NeuroImage* 21.4 (2004): 1748-1761.

World Health Organization (2020). “The top 10 causes of death,” 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ the-top-10-causes-of-death

Wu, Haijing, et al. (2020). “A deep learning, image based approach for automated diagnosis for inflammatory skin diseases.” *Annals of translational medicine* 8.9 (2020).

Wu, Yanan, et al. (2021). “A vision transformer for emphysema classification using CT images.” *Physics in Medicine & Biology* 66.24 (2021): 245016.

Yadav, Samir S., and Shivajirao M. Jadhav (2019). “Deep convolutional neural network based medical image classification for disease diagnosis.” *Journal of Big Data* 6.1 (2019): 1-18.

Yang, Jie, et al. (2016a). “Emphysema quantification on cardiac CT scans using hidden Markov measure field model: the MESA Lung Study.” *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2016.

Yang, Jie, et al. (2016b). “Explaining radiological emphysema subtypes with unsupervised texture prototypes: MESA COPD study.” *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging.* Springer, Cham, 2016. 69-80.

Yang, Jie, et al. (2017). “Unsupervised discovery of spatially-informed lung texture patterns for pulmonary emphysema: The MESA COPD study.” *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2017.

Yang, Jie, et al. (2019). “Unsupervised Domain Adaption with Adversarial Learning (UDAA) for Emphysema Subtyping on Cardiac CT Scans: The Mesa Study.” *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019.

Yang, Jie, et al. (2021). “Novel subtypes of pulmonary emphysema based on spatially-informed lung texture learning: the multi-ethnic study of atherosclerosis (MESA) COPD Study.” *IEEE transactions on medical imaging* 40.12 (2021): 3652-3662.

Ye, Hai, et al. (2019). “Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network.” *European radiology* 29.11 (2019): 6191-6201.

Zhang W et al. (2015). “Deep convolutional neural networks for multi-modality isointense infant brain image segmentation.” Neuroimage 108:214–224, 2015

Zhao, Zhong-Qiu, et al. (2019). “Object detection with deep learning: A review.” *IEEE transactions on neural networks and learning systems* 30.11 (2019): 3212-3232.

Zhou, Pan, et al. (2020). “Towards theoretically understanding why sgd generalizes better than adam in deep learning.” *arXiv preprint* arXiv:2010.05627 (2020).