

**Machine Learning Application to Study Human Brain: The
Investigation of Brain Microstructure and Speech Decoding
based on Cortical Neural Activity**

DISSERTATION

Submitted in Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY (Electrical Engineering)

at the

**NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING**

by

Junbo Chen

January 2024

Machine Learning Application to Study Human Brain: The Investigation of Brain Microstructure and Speech Decoding based on Cortical Neural Activity

DISSERTATION

Submitted in Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY (Electrical Engineering)

at the

**NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING**

by

Junbo Chen

January 2024

Approved:



Department Chair Signature

December 5, 2023

Date

University ID: N15094863

Net ID: jc7489

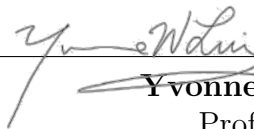
Approved by the Guidance Committee:

Major: Electrical Engineering



12/5/2023

Yao Wang
Professor
NYU Tandon School of Engineering



12/5/2023

Yvonne W. Lui
Professor
NYU Grossman School of Medicine



12/04/2023

Anna Choromanska
Assistant Professor
NYU Tandon School of Engineering

Microfilm or other copies of this dissertation are obtainable from

UMI Dissertation Publishing
ProQuest CSA
789 E. Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Vita

Junbo Chen was born in Tianjin, China in 1994. He received his Bachelor's degree in Automation from Wuhan University, Wuhan, Hubei, China in 2016. He then earned his Master of Science in Electrical Engineering from Columbia University, New York, NY, United States in 2017. He started his Ph.D. program in Electrical Engineering at New York University in 2019. His Ph.D. research focuses on computer vision, deep learning, machine learning, medical imaging and neural engineering. During his Ph.D. years, he also interned at TuSimple and Meta, worked on machine learning related research.

Acknowledgements

I am grateful to all the people who have helped me throughout my Ph.D. study.

First and foremost, I would like to thank my Ph.D. advisor Prof. Yao Wang for her unlimited support and guidance throughout the course of my Ph.D. study. Prof. Yao Wang is the best Ph.D. advisor I can ever imagine. I would also like to thank Prof. Yvonne W. Lui and Prof. Adeen Flinker for their generous guidance and advice throughout my research. I would like to thank Prof. Anna Choromanska for her suggestions and kindness. Furthermore, I would like to thank my collaborators Xupeng Chen, Dr. Ran Wang, Dr. Amirhossein Khalilian-Gourtani, Chenqian Le, Antoine Ratouchniak, Nika Emami, Tianhao Li, Vara Lakshmi Bayanagari, Prof. Els Fieremans, Prof. Dmitry S. Noviko, Prof. Sohae Chung for the assistances and helps in our research and people who have provided me with suggestions.

Last but not least, I would like to express sincere gratitude to my parents and my wife, Fan Duan, for their unlimited support.

Junbo Chen

January 2024

To all the Ph.D. pursuing brave souls

ABSTRACT

**Machine Learning Application to Study Human Brain: The
Investigation of Brain Microstructure and Speech Decoding based on
Cortical Neural Activity**

by

Junbo Chen

Advisor: Yao Wang, Ph.D.

**Submitted in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy (Electrical Engineering)**

January 2024

Machine learning and deep neural networks have succeeded in various computer vision tasks involving modalities ranging from natural to medical images. The advancement of machine learning can shed light on studying the human brain. This thesis leverages machine learning to study the human brain from two aspects: understanding the microstructure of the brain and designing a neural activity decoder to predict human speech.

For the investigation of brain microstructure, we leverage the multi-shell diffusion MRI to capture the brain structure at the microscopic level. We first investigate the microstructural alteration caused by repeated head impacts (RHI). We propose a classification and feature selection pipeline as a means towards identifying important diffusion metrics associated with RHI. The results support the notion that there are detectable white matter microstructure changes in the setting of RHI and pinpoint influential diffusion metrics. The work serves as an example of methods that lead to a better understanding of the myriad of diffusion metrics as they relate to injury and disease. We also investigate the sex-related (biological sex assigned at birth) differences at the microscopic level in the human brain by classifying sex with deep neural networks based on registered diffusion metrics, which can pave the way for understanding brain disorders that manifest differently in different sexes. The study designs 3D CNN, 2D CNN, and Vision Transformer sex classifiers based on multiple volumetric diffusion metrics to capture complementary information. Given models with promising accuracy, occlusion analysis is applied to determine which white matter regions contribute most to sex-related differences. The results provide new insight supporting differences between male and female brain cellular-level tissue.

For decoding speech from human neural activity, we design a novel neural network architecture to decode speech from electrocorticography (ECoG) recordings and a semi-supervised pretraining method for this ECoG decoder. We first propose a novel ECoG speech decoder, named SwinT. Instead of relying on any grid index, the SwinT leverages each electrode’s anatomical position and brain parcellation to decode human speech, enabling the model architecture to accommodate arbitrarily positioned electrodes. The proposed model achieved state-of-the-art performance based on the same grid electrodes used in the previous studies. It also achieved

further performance increases by leveraging off-grid electrodes. More importantly, instead of relying on subject-specific ECoG decoders, our SwinT can be trained with ECoG signals from multiple subjects. The SwinT trained with multiple subjects not only achieved performance increase but also demonstrated generalizability to unseen subjects outside of the training set. For subjects included during training, to further improve speech decoding, we propose a novel semi-supervised pretraining approach for feature extraction part of the SwinT decoder. The study aims to simplify the complex neural activity associated with speech production by decomposing the latent representation into word-level semantics and trial-level dynamics. The pretraining framework combines the pretasks of neural signal reconstruction and contrastive learning to guide the decomposition. Refining the pretrained network with the decoding loss led to improved speech decoding performance compared to training from scratch.

Table of Contents

Vita	iv
Acknowledgements	v
Abstract	vii
List of Figures	xxiv
List of Tables	xxv
1 Introduction	1
1.1 Overview	1
1.2 Problem Statement	2
1.3 Contributions	4
1.4 Organization of the Thesis	5
2 Identifying Relevant Diffusion MRI Microstructure Biomarkers Relating to Exposure to Repeated Head Impacts in Contact-Sport Athletes	6
2.1 Introduction	6
2.2 Methods	8
2.2.1 Feature Generation	8
2.2.2 Classification Pipeline	9

2.2.2.1	Feature Selection	9
2.2.2.2	Classifier and Feature Importance	10
2.2.2.3	Experimental Details, Metric and Brain Region Importance	11
2.2.3	Results	13
2.2.3.1	Diffusion MRI Acquisition and Processing	13
2.2.3.2	Classification Performance in the Context of Feature Selection	15
2.2.3.3	Diffusion Metric Importance in the Identification of Subconcussive RHI	15
2.2.3.4	Brain Region Importance in the Identification of Subconcussive RHI	16
2.2.4	Discussion	16
2.2.5	Conclusion and Contributions	22

3 Deep Learning with Diffusion MRI as in vivo Microscope Reveals

Sex-related Differences in Human White Matter Microstructure 23

3.1	Introduction	23
3.2	Methods	27
3.2.1	End-to-End Classification Models	27
3.2.1.1	2D Convolutional Neural Network	27
3.2.1.2	3D Convolutional Neural Network	30
3.2.1.3	Vision Transformer for 3D Input Pretrained with Mask Autoencoders	31
3.2.2	Model Training and Evaluation	33
3.2.3	Occlusion Analysis	35

3.3	Result	35
3.3.1	Diffusion MRI Acquisition and Processing	35
3.3.2	Classification Results	37
3.3.3	Occlusion Analysis Results	38
3.4	Discussion	38
3.5	Conclusion and Contributions	45
4	Temporal Swin Transformer for Grid-Free ECoG Speech Decoding on Single and Multi Patient	47
4.1	Introduction	47
4.2	Method	49
4.2.1	Speech Decoding Framework	49
4.2.2	Grid-Free ECoG Decoder based on Temporal Swin Transformer	52
4.2.3	Multi-Subject ECoG-to-Audio	57
4.2.4	Training of Speech Encoder and Speech Synthesizer	59
4.2.5	Training of ECoG Decoder	60
4.3	Results	61
4.3.1	ECoG Data Collection and Preprocessing	61
4.3.2	Single-Subject Speech Decoding with Grid Electrodes	63
4.3.3	Speech Decoding with Additional Off-Grid Electrodes	63
4.3.4	Speech Decoding Trained with Multiple-Subjects	66
4.4	Discussion	70
4.5	Conclusion and Contributions	74
5	Semi-Supervised Learning for ECoG Decoder based on Latent Decomposition	75

5.1	Introduction	75
5.2	Method	77
5.2.1	Pretraining Framework with Latent Decomposition	77
5.2.2	Data Augmentation	80
5.2.3	Loss Functions and Training	81
5.3	Results	83
5.4	Discussion	84
5.5	Conclusion and Contributions	86
6	Conclusion	87

List of Figures

- 2.1 Seven major WM regions of interest (ROI) used in this study to calculate diffusion features: The corpus callosum plus supratentorial hemispheric WM divided into rostral, middle, and caudal hemispheric segments divided based on boundaries from John Hopkins University (JHU) ICBM-DTI-81 WM atlas. 9
- 2.2 Schematic of the experimental pipeline: we conduct wrapper-based feature selection using training data for each classifier, with average AUC from 10 random repetitions of stratified 5-fold cross-validation as the validation performance. Hyperparameters are tuned based on cross-validation using the training set; for each classifier type, feature subset and hyperparameters with best average cross-validation AUC are selected; each classifier is trained using the entire training set and finally tested on the held-out test set. 13
- 2.3 Relative importance of selected diffusion metrics in identifying RHI-related diffusion changes, derived from the sum of feature importance scores for each diffusion metric across all regions-of-interest and all statistics: (a) logistics regression; (b) linear SVM. 17

2.4 Relative importance of ROIs in identifying RHI-related diffusion changes; derived from the sum of the importance scores for each ROI across all diffusion metrics and all statistics: (a) logistic regression; (b) linear SVM. The relative importance of the various ROIs is not consistent between the two classifiers. 18

3.1 Our 2D CNN model. In the top of the figure, the imaging volume is divided into subvolumes, and a shared ResNet18 is applied to extract 512 features from each subvolume. The features are concatenated and fed to a linear layer for the final prediction. The bottom of the figure shows the architecture of ResNet18 (residual connection, ReLU activation, batch normalization are omitted for simplicity): The input is first fed into a convolutional layer (7x7 kernel-size, stride=2, channel-number=64) followed by a max-pooling (kernel-size=3x3, stride=2) layer; subsequently, 8 residual blocks are applied with each containing 2 convolutional layers. Residual blocks parameters: conv layers in block 1, 2 have kernel-size=3x3 and channel=64; conv layers in block 3, 4 have kernel-size=3x3 and channel=128; conv layers in block 5, 6 have kernel-size=3x3 and channel=256; conv layers in block 7, 8 have kernel-size=3x3 and channel=512; stride=2 is applied at the first conv layer of block 3, 5, 7. Global average pooling is applied at the end. 29

- 3.2 Our 3D CNN model based on ResNet10 (residual connection, ReLU activation, group normalization omitted for simplicity). The 3D volume is first fed to a conv layer (kernel-size=7x7x7, stride=2, channel=64) followed by a max pooling (kernel-size=3x3x3, stride=2). Subsequently, 8 residual blocks are applied with each containing 1 conv layer. Residual blocks parameters: block 1, 2 have kernel-size=3x3x3 and channel=64; block 3, 4 have kernel-size=3x3x3 and channel=128; block 5, 6 have kernel-size=3x3x3 and channel=256; block 7, 8 have kernel-size=3x3x3 and channel=512; stride=2 is used at conv layer in block 3, while dilation=2 is used at conv layer in block 5 and dilation=4 is used at conv layer in block 7. Global average pooling is applied at the end. 30
- 3.3 Vision Transformer for Diffusion MRI sex classification: the imaging volume inputted is partitioned into non-overlapping patches. Each patch is projected to patch embedding using a linear patch embedding layer, and added with positional embedding representing the position of the patch. A classification token is appended to the sequence of tokens to learn representation of the entire input sample. The structure of the transformer encoder is shown on the right, which consists of L alternating layers of multi-head attention and multiple-linear-perceptron (MLP) blocks. After the transformer encoder, the corresponding output of the classification token is fed to the classification head to generate prediction results. 32

3.4	WM regions in selected slices with significant ($p < 0.05$) impact on classification probability based on occlusion analysis for FA; numbered labels based on JHU-ICBM-1mm atlas (https://identifiers.org/neurovault.image:1401); 1: middle cerebellar peduncle, 2: pontine crossing tract (a part of middle cerebellar peduncle), 3: genu of corpus callosum, 4: body of corpus callosum, 5: splenium of corpus callosum, 7: corticospinal tract (right), 9: medial lemniscus (right), 10: medial lemniscus (left), 14: superior cerebellar peduncle (left); 17: anterior limb of internal capsule (right), 20: posterior limb of internal capsule (left), 35: cingulum (cingulate gyrus) (right), 37: cingulum (hippocampus) (right), 40: stria terminalis (left), 48: tapetum (left).	39
-----	--	----

3.5	WM regions in selected slices with significant ($p < 0.05$) impact on classification probability based on occlusion analysis for MD; numbered labels based on JHU-ICBM-1mm atlas (https://identifiers.org/neurovault.image:1401); 1: middle cerebellar peduncle, 2: pontine crossing tract (a part of middle cerebellar peduncle), 3: genu of corpus callosum, 4: body of corpus callosum, 5: splenium of corpus callosum, 5: splenium of corpus callosum, 7: corticospinal tract (right), 13: superior cerebellar peduncle (right), 14: superior cerebellar peduncle (left), 15: cerebral peduncle (right), 17: anterior limb of internal capsule (right), 18: anterior limb of internal capsule (left), 19: posterior limb of internal capsule (right), 20: posterior limb of internal capsule (left), 22: retrolenticular part of internal capsule (left), 25: superior corona radiata (right), 26: superior corona radiata (left), 27: posterior corona radiata (right), 28: posterior corona radiata (left), 31: sagittal stratum (right), 35: Cingulum (cingulate gyrus) (right), 36: cingulum (cingulate gyrus) (left), 37: cingulum (hippocampus) (right), 39: stria terminalis (right), 40: Stria terminalis (left), 42: superior longitudinal fasciculus (left), 43: superior fronto-occipital fasciculus (right).	40
-----	--	----

3.6 WM regions in selected slices with significant ($p < 0.05$) impact on classification probability based on occlusion analysis for MK; numbered labels based on JHU-ICBM-1mm atlas (<https://identifiers.org/neurovault.image:1401>); 1: middle cerebellar peduncle, 2: pontine crossing tract (a part of middle cerebellar peduncle), 4: body of corpus callosum, 5: splenium of corpus callosum, 6: fornix (column and body of fornix), 26: superior corona radiata (left), 37: Cingulum (hippocampus) (right), 38: cingulum (hippocampus) (left), 46: uncinate fasciculus (left). 41

4.1 2-Step Speech Decoding Training Framework (step1) Audio-to-Audio Training: the original speech waveform is first converted to speech spectrogram with STFT, then speech parameters are generated at each frame based on speech spectrogram by Speech Encoder. A speech synthesizer is trained to reconstruct the speech spectrogram based on speech parameters from the Speech encoder. (step2) ECoG-to-Audio Training: The ECoG decoder maps ECoG high-gamma signal to latent representation and predicts speech parameters supervised by the speech parameters from the trained speech encoder from step 1. The predicted speech parameters from the ECoG decoder are fed to the trained speech synthesizer from step 1 to generate the predicted speech spectrogram, which is reversed to the predicted speech signal. 51

4.2 **a.** SwinT ECoG decoder. SwinT uses three stages of temporal swin transformer blocks with spatial-temporal attention with temporal windowing to extract features. An MLP layer is applied to decrease latent dimension. Spatial max pooling is then applied followed by transposed temporal convolution to upsample the temporal dimension. **b.** Prediction head for speech parameters consists of temporal convolution and MLP that map features from (a) to speech parameters at every frame. 56

4.3 multiple-subject ECoG decoding training pipeline. Given multiple subjects, each subject’s ECoG signal and electrodes’ location information (MNI coordinates and ROI region index) are fed to a shared SwinT ECoG decoder to predict speech parameters. The predicted speech parameters are supervised by the speech parameters generated by subject-specific speech encoder from ground-truth speech spectrogram. The each subject’s predicted speech parameters are then fed into the corresponding subject-specific speech synthesizer to generate speech spectrogram. 58

4.4 ECoG electrodes implanted on the human brain (a) 8x8 macro electrodes on the grid (b) both grid and off-grid electrodes. 62

4.5 Comparison between baseline ECoG decoder and proposed SwinT when both trained and tested on grid electrodes: **(a)**: Comparison between ResNet and SwinT regarding PCC; **(b)**: Comparison between 3D Swin and SwinT regarding PCC; **(c)**: Comparison between ResNet and SwinT regarding STOI+; **(d)**: Comparison between 3D Swin and SwinT regarding STOI+. Each point indicates the speech decoding performance of a specific subject, with the x-axis as the performance of the SwinT and the y-axis as the performance of the baseline model. For (a)-(d), all sample points are below the diagonal line with only a few exceptions, indicating that the SwinT outperforms the two baseline models regarding both PCC and STOI+. 64

4.6 Comparison between baseline ECoG decoders (based on ResNet and 3D Swin Transformer) and the proposed SwinT ECoG decoder when models are trained and tested with grid electrodes for each subject individually: **(a)**: Comparison regarding PCC; **(b)**: Comparison regarding STOI+. The SwinT outperforms the baseline ECoG decoders based on ResNet and 3D Swin Transformer regarding both PCC and STOI+. 65

- 4.7 Comparison between SwinT ECoG decoders w and w/o off-grid electrodes, when trained on each subject individually. **(a) and (b)**: Comparison in PCC and STOI+, respectively. Each point indicates a subject, with x-axis being the performance of the SwinT with off-grid electrodes and y-axis being the performance of the Swin without off-grid electrodes; **(c) and (d)**: Comparison in PCC and STOI+ box plot. The results indicate the superior performance of the SwinT with off-grid electrodes. 67
- 4.8 Comparison between the SwinT ECoG decoder trained with multiple (15) subject and the subject-specific SwinT. PCC and STOI+ were evaluated on test trials from the 15 subjects. 68
- 4.9 The performance of SwinT inferred on unseen subjects that are not in the training set. **(a)** cross-validation conducted on male subjects; **(b)** cross-validation conducted on female subjects. For each plot, the speech decoding performance when subjects are outside of the training subjects is shown on the right, and the performance of the SwinT decoder trained on each specific subject is shown on the left. The results demonstrate that the SwinT ECoG decoder can achieve generalizability to unseen subjects, as the performance achieved on unseen subjects is in a range with significant overlap with the performance of the single-subject models. 69

4.10	The comparison of speech decoding performance on unseen subjects between SwinT trained on one hemisphere and SwinT trained on both hemispheres. Models were trained separately for males and females. The results demonstrate that, compared with hemisphere-specific models, the SwinT ECoG decoder trained on both hemispheres can achieve comparable or slightly better performance when inferenced on unseen subjects.	70
5.1	Framework for the ECoG decoder pretraining. At each iteration, three samples are input to the framework: two trials from the same word and one trial from a different word. The SwinT encoder extracts latent representation from each sample. The representation is divided into two parts with same dimension: word latent and trial latent, denoated as W and T . The latents of the original three latents go through decoder to reconstruct the corresponding signals. Besides, the content of the two same-word trials are swapped and used to predict the signals corresponding to the trial latent. The word latent are also fed to word classifier with contrastive loss.	78
5.2	(a) : SwinT Encoder, detailed in Section 4.2.2; (b) : decoder consists of four transposed conv layers with kernel-size=3x1 to predict the signal reconstruction. $C = 96$, $C' = 64$, $C'' = 128$	79
5.3	(a) : Word classification prediction head; (b) : If-same-word classification prediction head.	80

5.4	Performance comparsion between ECoG decoder with and without pretraining. The performance of speech decoding are measured as PCC and STOI+. For pretrained SwinT ECoG decoder, we evaluated the speech decoding performances that rely on word+trial latent, word latent only, and trial latent only.	84
-----	---	----

List of Tables

2.1	Study Cohort	14
2.2	Test AUC of 5 different classification models using selected features based on regional ROIs, WM skeleton, and whole WM. (The numbers in parenthesis are the mean and standard deviation of AUC among validation folds)	16
3.1	Study Cohort	36
3.2	Performance (test AUC) of sex classification models using three different diffusion MRI parametric maps as inputs (FA, MD, and MK)	37
3.3	Number of white matter regions showing significant differences between males and females in the occlusion analysis; 48 WM regions in total.	38

Chapter 1

Introduction

1.1 Overview

The machine learning and deep neural networks have achieved successes in various computer vision studies involving modalities from natural [39, 57, 58, 77, 78, 79] to medical images [52, 59, 61, 131]. The advancement of machine learning can shed light on studying the human brain. This thesis leverages machine learning to study the human brain from two aspects: understanding the microstructure of the brain and designing a neural activity decoder to predict human speech.

For the investigation of brain microstructure, we study the microstructural alteration caused by repeated head impacts (RHI), and the sex-related (biological sex assigned at birth) differences in the human brain at the microscopic level. We leverage the multi-shell diffusion MRI to capture the microstructure of brain tissue. In the study of RHI, we propose a classification and feature selection pipeline as a means towards identifying important diffusion metrics associated with the RHI. In the study of sex-related differences, we design 2D convolutional neural networks

(CNN) [58], 3D CNN [22], and Vision Transformer (ViT) [39] sex classifiers based on multiple volumetric diffusion metrics to capture complementary information. The ViT is pretrained with masked auto-encoding task [57]. Given models with promising accuracy, occlusion analysis is applied to determine which brain regions contribute most to sex-related differences.

For decoding speech from human neural activity, we design a novel neural network architecture to decode speech from electrocorticography (ECoG) recordings and a semi-supervised pretraining method for this ECoG decoder. We first propose a novel ECoG speech decoder, named SwinT. Instead of relying on any grid index, the SwinT leverages each electrode’s anatomical position and brain parcellation to decode human speech, enabling the model architecture to accommodate arbitrarily positioned electrodes. The proposed SwinT can leverage off-grid electrodes. Besides, the SwinT can be trained with ECoG signals from multiple subjects and achieve generalizability to unseen subjects outside of the training set. To further improve speech decoding with limited training data, we propose a novel semi-supervised pretraining approach for the feature extraction part of the SwinT decoder. The pretraining method aims to simplify the complex neural activity associated with speech production by decomposing the latent representation into word-level semantics and trial-level dynamics. The pretraining framework combines the pretasks of neural signal reconstruction and contrastive learning to guide the decomposition.

1.2 Problem Statement

For the investigation of brain microstructure, we focus on leveraging classification as means of identifying microstructural differences of the target cohorts. Specifically,

we design models to classify the target cohorts based on diffusion MRI. Given classifiers achieving promising classification performance, we interpret the models by analyzing the learned weights or conducting occlusion analysis to pinpoint important diffusion metrics or brain regions for the classification tasks. The findings are then used to provide new insights into the microstructural differences of the target cohorts. The study of microstructure alteration associated with RHI has the challenges of limited data, the subtlety of microstructural differences, and insufficient prior knowledge. We solve the challenges by designing a classification pipeline with hand-crafted features and wrapper-based feature selection to distinguish RHI subjects, with the learned weights of promising models being analysed for clinical insights. The study of sex-related differences also has the challenges of limited data and insufficient prior knowledge, with findings not entirely consistent across previous studies. In our study, we leverage multiple distinctive neural network architectures to capture complementary sex-related differences, with self-supervised pretraining applied to the data-demanding classifier. Occlusion analysis is applied for clinical insights.

For decoding speech from human neural activity, we design ECoG decoder that can predict speech parameters at every time frame from ECoG recordings collected from patients. The speech signals generated from the predicted speech parameters are evaluated and compared with the ground truth. The study has many challenges: the dataset is limited, the layout of ECoG electrodes does not follow grid-topology, and the placement of electrodes has differences among subjects. To solve these challenges, we design a grid-free ECoG decoder to leverage electrodes that can not fit into a grid and leverage ECoG signals from multiple subjects. As the dataset is limited, we also design a semi-supervised pretraining method consisting of neural

signal reconstruction and contrastive learning to improve the speech decoding performance of subjects in the training set.

1.3 Contributions

In the study of microstructural differences of RHI, the proposed classification and feature selection pipeline achieved promising results in classifying RHI subjects. The results support the notion that there are detectable white matter microstructural changes in the setting of RHI due to playing contact sports. The learned weights of classifiers are used to pinpoint influential diffusion metrics associated with RHI. The work serves as an example of methods that leveraging machine learning to gain a better understanding of the myriad of diffusion metrics as they relate to injury and disease.

In the study of sex-related differences at the microscopic level, the designed 3D CNN, 2D CNN, and Vision Transformer sex classifiers can achieve promising sex classification performance based on multiple volumetric diffusion metrics. Occlusion analysis is applied to determine which white matter regions contribute most to sex-related differences. The results indicate that distinctive neural networks can capture complementary information regarding sex-related differences. And the results provide new insight supporting differences between male and female brain cellular-level tissue.

For decoding speech from human neural activity, we first propose a novel ECoG speech decoder, named SwinT. Instead of relying on any grid index, the SwinT leverages each electrode’s anatomical position and brain parcellation to decode human speech, enabling the model architecture to accommodate arbitrarily

positioned electrodes. The proposed model achieved state-of-the-art performance based on the same grid electrodes used in the previous studies. It also achieved further performance increases by leveraging off-grid electrodes. More importantly, instead of relying on subject-specific ECoG decoders, our SwinT can be trained with ECoG signals from multiple subjects. The SwinT trained with multiple subjects not only achieved performance increase but also demonstrated generalizability to unseen subjects outside of the training set.

We further propose a novel semi-supervised learning method to pretrain the SwinT with selected pretasks: neural signal reconstruction, contrastive learning, and word classification. Refining the pretrained network with the decoding loss is shown to lead to improved speech decoding performance compared to training from scratch.

1.4 Organization of the Thesis

In the following chapters, we first introduce our studies about the brain microstructure: the study of RHI is introduced in Chapter 2, and the study of sex-related microstructural differences is introduced in Chapter 3. We then introduce the studies of speech decoding based on ECoG recordings. In Chapter 4, we introduce our non-grid ECoG speech decoder SwinT. In Chapter 5, we introduce the proposed semi-supervised learning method to pretrain the SwinT to improve speech decoding performance. We finally summarize our works in Chapter 6.

Chapter 2

Identifying Relevant Diffusion

MRI Microstructure Biomarkers

Relating to Exposure to Repeated

Head Impacts in Contact-Sport

Athletes

2.1 Introduction

Recently, exposure to repeated head impacts (RHI) due to playing contact sports has emerged as a potential health concern [107]. This is true even in the absence of frank concussion. RHI exposure sustained over a long period is associated

Junbo Chen is the main driver of this study. Acknowledgment to Prof. Sohae Chung, Tianhao Li, Prof. Els Fieremans, Prof. Dmitry S. Novikov, Prof. Yao Wang, and Prof. Yvonne W. Lui for their collaboration and advice.

with negative downstream effects on cognition [114] as well as increased risk of neurodegenerative disorders including movement disorders such as Parkinson’s disease [80] and behavioral disorders such as chronic traumatic encephalopathy [11]. Diffusion MRI has been used to study the in vivo changes to brain microstructure after RHI [7, 12, 21, 30, 35, 46, 86, 87, 107]. Published studies rely almost uniformly on conventional statistical methods to analyze group-level differences of individual diffusion metrics separately such as fractional anisotropy (FA) [7, 12, 21, 30, 35, 46, 86, 107], mean diffusivity (MD) [12, 21, 46, 86, 107], and mean kurtosis (MK) [30, 35]. With advances in non-Gaussian approaches and compartment modeling of diffusion signal, our ability to characterize biophysical characteristics of white matter has progressed; however, as a result, diffusion MRI has become ever more challenging to interpret because of the sheer variety and number of metrics (both empiric and modeled) that can be used. There is a growing need to hone in on which of many metrics are really relevant to disease and injury. In this work, we employ a classification task not as an end but as a means to highlighting the most relevant diffusion MRI metrics to try to better understand the pathophysiology of RHI and to provide a proof-of-concept method towards parsing multidimensional diffusion data in a limited study cohort.

The purpose of this study is to investigate white matter (WM) microstructure in collegiate contact sport athletes exposed to subconcussive RHI by identifying diffusion metrics across a combination of diffusion methods that are most useful in discriminating between athletes with exposure to RHI and non-contact sport controls. We include metrics from standard empiric methods of diffusion MRI (diffusion tensor imaging (DTI), diffusion kurtosis imaging (DKI)) as well as modeled metrics which have become increasingly popular to uncover biophysically meaningful

diffusion information, specifically the two-compartment model of diffusion signal (white matter tract integrity (WMTI)). In this work, we develop the classifier not as a goal itself but instead to identify the most relevant diffusion MRI biomarkers that may be important in detecting/quantifying microstructural changes due to RHI exposure.

2.2 Methods

2.2.1 Feature Generation

We study both global and regional WM features. Global WM features were calculated based on the standard FA template (Montreal Neurological Institute 152 space) [85] using 2 methods: whole brain WM volume as well as WM skeleton derived from the FSL Tract-based Spatial Statistics analysis (TBSS) [110]. For regional analysis, 7 WM regions were defined and used to generate regional features (corpus callosum, supratentorial hemispheric WM divided into rostral, middle, and caudal regions based on the John Hopkins University ICBM-DTI-81 WM atlas (shown in Figure 2.1) [89]). Each subject’s FA map was registered to the standard FA template in FMRIB Software Library (FSL) [111], followed by a reversed warping process to generate atlas labels in each unique subject space.

Diffusion maps were thresholded at FA of 0.4 to restrict the analysis to WM regions consisting primarily of single-fiber orientations as has been previously recommended for WMTI metrics [31, 43], the same threshold was applied to DTI and DKI metrics to make all metrics have consistent regions. For regional features, to compactly represent the statistical distribution of diffusion metrics within regions, 3 basic statistical features (mean, standard deviation and skewness) are generated

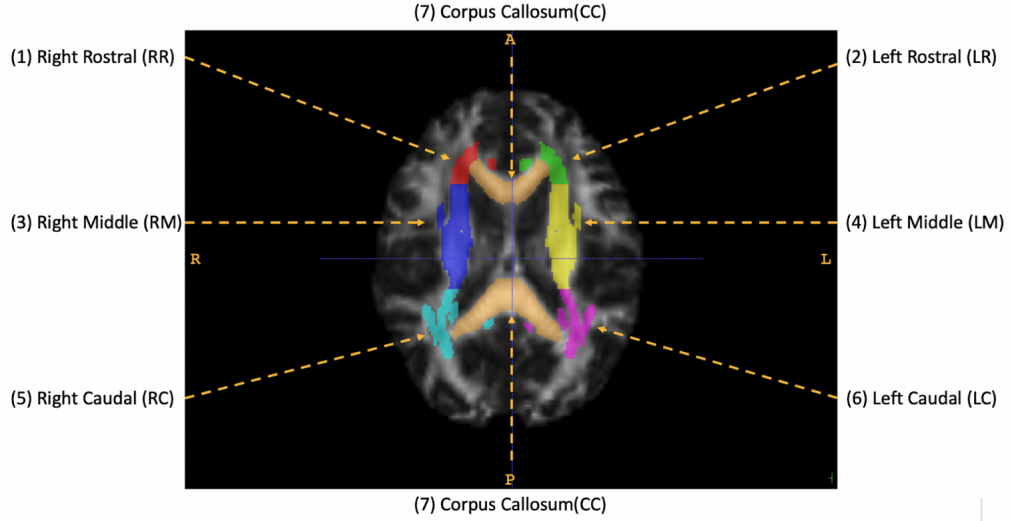


Figure 2.1: Seven major WM regions of interest (ROI) used in this study to calculate diffusion features: The corpus callosum plus supratentorial hemispheric WM divided into rostral, middle, and caudal hemispheric segments divided based on boundaries from John Hopkins University (JHU) ICBM-DTI-81 WM atlas.

for each region and each diffusion metric, yielding a total of 147 features for each scan. For the 2 types of global WM features, the same 3 statistics were calculated for each diffusion metric, yielding 21 features in each case.

2.2.2 Classification Pipeline

2.2.2.1 Feature Selection

As the dataset is relatively small and each sample has a multiple-dimensional array of diffusion metrics, feature selection was applied to address any potential issues of overfitting. A wrapper-based method was selected as a high performance [127] and flexible feature selection method that generally is adaptable to any classifier. For each candidate feature subset, a classifier (see next section for details on classifiers) was trained and performance on validation samples was used as

the goodness metric for this feature subset. The best feature subset for each classifier was derived by cycling through possible feature subsets and a greedy search [72] with a crossover operator [127] was used when the number of feature combinations was too large to be efficiently traversed exhaustively. Using this approach enabled us to identify the most promising feature subset based on cross-validation performance ending when the performance was not improved upon over 100 iterations. Furthermore, the crossover operator inspired from the genetic algorithm [74] was employed, effectively and efficiently generating promising features at a low computational cost [127].

2.2.2.2 Classifier and Feature Importance

Five classifiers are tested in this study. These classifiers are selected based on 1) interpretability with embedded ability to identify influential diffusion metrics (the purpose of the study), 2) robustness against overfitting given relatively small dataset, and 3) the array of 5 classifiers included represent the 5 most widely used types of classification algorithms: logistic regression [16], linear support vector machine (SVM) [16], SVM with radial basis function (RBF) kernel [16], gradient boosting trees (GBT) [26, 44] and multi-layer perceptron (MLP) [55]. SVM with nonlinear kernel, GBT and MLP are all nonlinear approaches and have higher learning capacity, but also a tendency to overfit in the case of limited data. These five approaches to classification provide a good representation of the range of learning capacities across classic classifiers to help us define the optimal trade-off between learning capacity and generalizability for this particular question.

As noted, all 5 of these classifiers benefit from good interpretability and feature importance can be derived from them. With logistic regression and linear SVM, the

weight associated with each feature directly reflects the feature’s importance. GBT is an ensemble model by boosting decision trees and provides feature importance based on average performance gain from each feature among all the trees. In the MLP classifier, feature importance can be inferred from the gradient of the classifier output to the input using a process similar to guided-backpropagation [112].

2.2.2.3 Experimental Details, Metric and Brain Region Importance

For each cohort, 28 studies from 7 randomly selected players (4 separate scans per individual collected at different visits throughout the season) were held out as unseen data for testing with balanced class distribution. All remaining data were used for classifier training and validation. Subject-wise data split was applied to prevent data leakage. Statistical features of all samples were preprocessed with Z-score normalization based on mean and standard deviation of each feature calculated from the training set. Area under the receiver operating characteristic curve (AUC) for the learned classifier are reported as an overall measure of reliability and predictive power; metrics such as classification accuracy, sensitivity and specificity are not reported here as the purpose of the classification task in this project is for identification of relative feature importance and thus there was no need to select thresholds on the predicted probability of each class.

An overall schematic of the experimental pipeline can be found in Figure 2.2. We use average AUC from 10 random repetitions of stratified 5-fold cross-validation as the validation performance. For each repetition, the training set is split randomly into 5 non-overlapping folds subject-wisely, with the same distribution of labels among subjects in each fold, and each fold is used as the validation set once while the other 4 folds are used for training. Such subject-wise stratified 5-fold cross-

validation is repeated 10 times and the averaged AUC from these 50 experiments is then used to evaluate the goodness of each feature subset. The hyperparameters of classifiers are tuned based on cross-validation using the training set as well. For MLP, we find that using two layers with 8 hidden nodes and ReLU activation has the best cross validation AUC, with Adam optimizer ($\beta_1=0.9$, $\beta_2=0.999$), learning rate as 0.001 and weight decay as 0.0001. Maximum depth as 2 and estimator number as 100 are used for GBT. The regularization term $C=0.1$ and $C=0.5$ are used for logistic regression and linear SVM respectively. For SVM with the RBF kernel, $C=0.1$ is used. The hyperparameter tuning is done outside the feature selection loop. That is, for each candidate hyperparameter setting, the best feature subset is chosen using the above wrapper-based feature selection method, with its corresponding goodness measure. The Scikit-learn package is used to implement the included classifiers [95]. For the top performing classifiers in the test set (test $AUC \geq 0.8$), we evaluate feature importance to help identify influential diffusion metrics and brain regions towards the characterization of WM alteration relating to subconcussive RHI. Given the feature importances derived from the top performing classifiers, rankings of diffusion metrics are derived by summing the importance of all chosen features associated with each diffusion metric. Similarly, importance of each ROI is evaluated by summing over all selected features associated with this ROI.

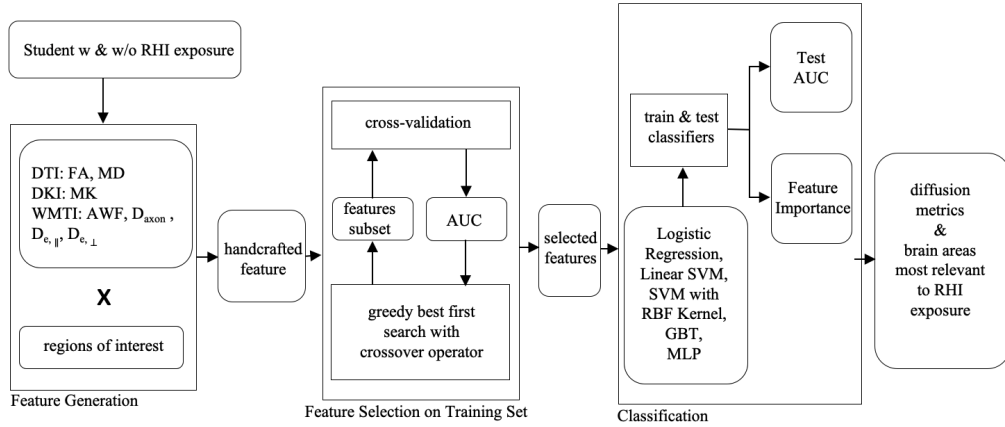


Figure 2.2: Schematic of the experimental pipeline: we conduct wrapper-based feature selection using training data for each classifier, with average AUC from 10 random repetitions of stratified 5-fold cross-validation as the validation performance. Hyperparameters are tuned based on cross-validation using the training set; for each classifier type, feature subset and hyperparameters with best average cross-validation AUC are selected; each classifier is trained using the entire training set and finally tested on the held-out test set.

2.2.3 Results

2.2.3.1 Diffusion MRI Acquisition and Processing

The study includes 125 diffusion images from 36 contact sport athletes (sub-concussive RHI subjects with history of head impact exposure) and 153 diffusion images from 45 non-contact sport controls (control subjects without a history of head impact exposure) from National Collegiate Athletic Association-Department of Defense Concussion Assessment, Research and Education (CARE) Consortium dataset [19], obtained over the course of three competitive collegiate athletic seasons. These data are available through the Federal Interagency Traumatic Brain Injury Research registry (FITBIR) and are part of the CARE study. Institutional review board approval and participants' informed consent were obtained at the

participating institutions. Participants were scanned up to 4 times throughout each season. Inclusion criteria for this study are: male sex, multi-shell diffusion MRI performed using 3T Prisma scanner (Siemens Medical Solutions, Erlangen, Germany), no history of concussion throughout the relevant seasons. Control subjects do not have a history of head impact exposure. Classification of contact versus non-contact sport followed the classification from the primary study and separate subjects based on the exposure to RHI. Contact sports include football, soccer, lacrosse. Non-contact control sports include baseball, cross country, track and field, basketball. Individuals with documented concussion during the study period were excluded. Demographic details are summarized in Table 2.1.

Table 2.1: Study Cohort

	Non-contact sport controls	Subconcussive RHI
subject number	45	36
scan number	153	125
Sex (M/F)	45/0	36/0
Age	19.9±1.3	19.6±1.2
Sport: subject (scan)		
Baseball	26 (85)	
Cross Country	15 (56)	
Tract and Field	3 (9)	
Basketball	1 (3)	
Football		29 (101)
Soccer		6 (20)
Lacrosse		1 (4)

Specifics of the diffusion MRI acquisition have been previously detailed in Broglio et al [19]. In brief, multi-shell diffusion images used the following parameters: 2 b-values (1000, 2000 s/mm²), 30 diffusion directions, 8 b₀ (b-value = 0) images, 2.7 mm isotropic image resolution, field of view = 243 mm x 243 mm, acquisition

matrix = 90 x 90, number of slices = 64, TR/TE = 7900/98 ms.

Preprocessing of the diffusion data includes Marchenko-Pastur principal component analysis based denoising [118], Gibbs correction [69], eddy current distortion, motion correction and outlier detection [33]. DESIGNER is used as the image-processing pipeline to preprocess diffusion data and generate diffusion metrics (DTI, DKI and WMTI metrics) as it demonstrated improved preprocessing accuracy compared to other processing methods [2]. We include 7 representative diffusion metrics: 2 DTI metrics (FA, MD), 1 DKI metric (MK), and 4 WMTI metrics (AWF, D_{axon} , $D_{e,\parallel}$ and $D_{e,\perp}$).

2.2.3.2 Classification Performance in the Context of Feature Selection

With region-based features, logistic regression achieved highest test AUC at 0.81 and second highest mean validation AUC at 0.83, whereas linear SVM achieved slightly lower test AUC at 0.80 and highest mean validation AUC at 0.86. The other 3 classifiers had substantially lower test AUC (0.48-0.67) as well as validation AUC (0.70-0.79). Use of the regional features showed substantially better performance over both whole WM-based features across all classifiers except for GBT. Details are summarized in Table 2.2.

2.2.3.3 Diffusion Metric Importance in the Identification of Subconcussive RHI

For the best performing 2 classifiers that achieved test AUC at 0.80-0.81 (logistic regression and linear SVM trained on region-based features), MD and MK are identified as the top 2 most important diffusion metrics across all regions for both classifiers (Figure 2.3), followed by the 4 WMTI metrics in the following order:

Table 2.2: Test AUC of 5 different classification models using selected features based on regional ROIs, WM skeleton, and whole WM. (The numbers in parenthesis are the mean and standard deviation of AUC among validation folds)

Classifiers	Regional ROIs	WM Skeleton	Whole Brain WM
Logistic Regression	0.81 (0.83±0.08)	0.63 (0.74±0.09)	0.81 (0.76±0.10)
Linear SVM	0.80 (0.86±0.07)	0.62 (0.75±0.10)	0.51 (0.75±0.10)
MLP	0.71 (0.79±0.08)	0.59 (0.77±0.10)	0.48 (0.79±0.08)
SVM RBF Kernel	0.64 (0.77±0.09)	0.59 (0.70±0.11)	0.49 (0.74±0.10)
GBT	0.64 (0.79±0.07)	0.67 (0.77±0.11)	0.58 (0.79±0.08)

$D_{e,\perp}$, $D_{e,\parallel}$, D_{axon} , AWF.

2.2.3.4 Brain Region Importance in the Identification of Subconcussive RHI

The best performing 2 classifiers on the test set (logistic regression and linear SVM trained on region-based features) lack clear consistency in ranking relevant importance of the different ROIs (Figure 2.4).

2.2.4 Discussion

In this work, by finding the relative importances of an array of diffusion MRI metrics for the classification of individuals with and without exposure to RHI, we are able to identify the metrics that may be most relevant to RHI. The two best performing classifiers show relative consistency in identifying the most influential diffusion metrics (Figure 2). Namely, MD and MK are considered the top two most important features by both these classifiers. MD represents the mean diffusivity in the area and MK is a marker of brain microstructure complexity. Overall, the current findings are in keeping with previous reports which reveal differences in MD

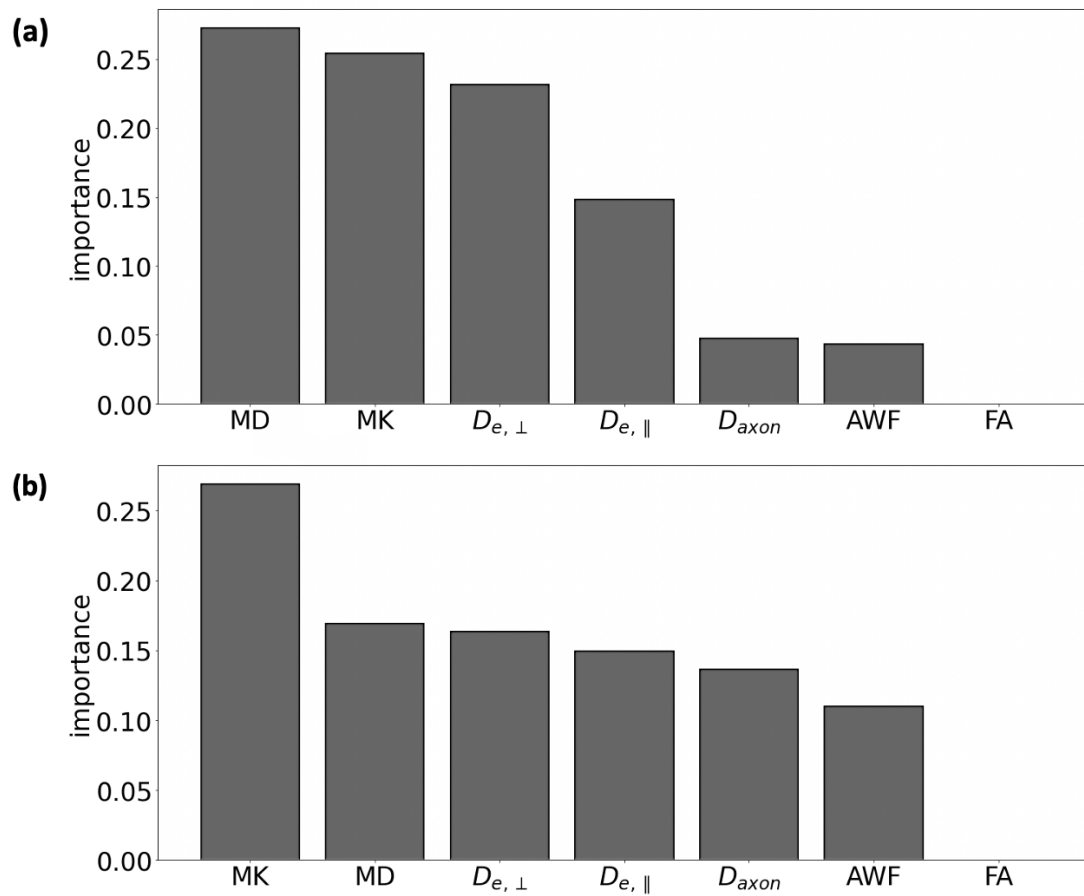


Figure 2.3: Relative importance of selected diffusion metrics in identifying RHI-related diffusion changes, derived from the sum of feature importance scores for each diffusion metric across all regions-of-interest and all statistics: (a) logistics regression; (b) linear SVM.

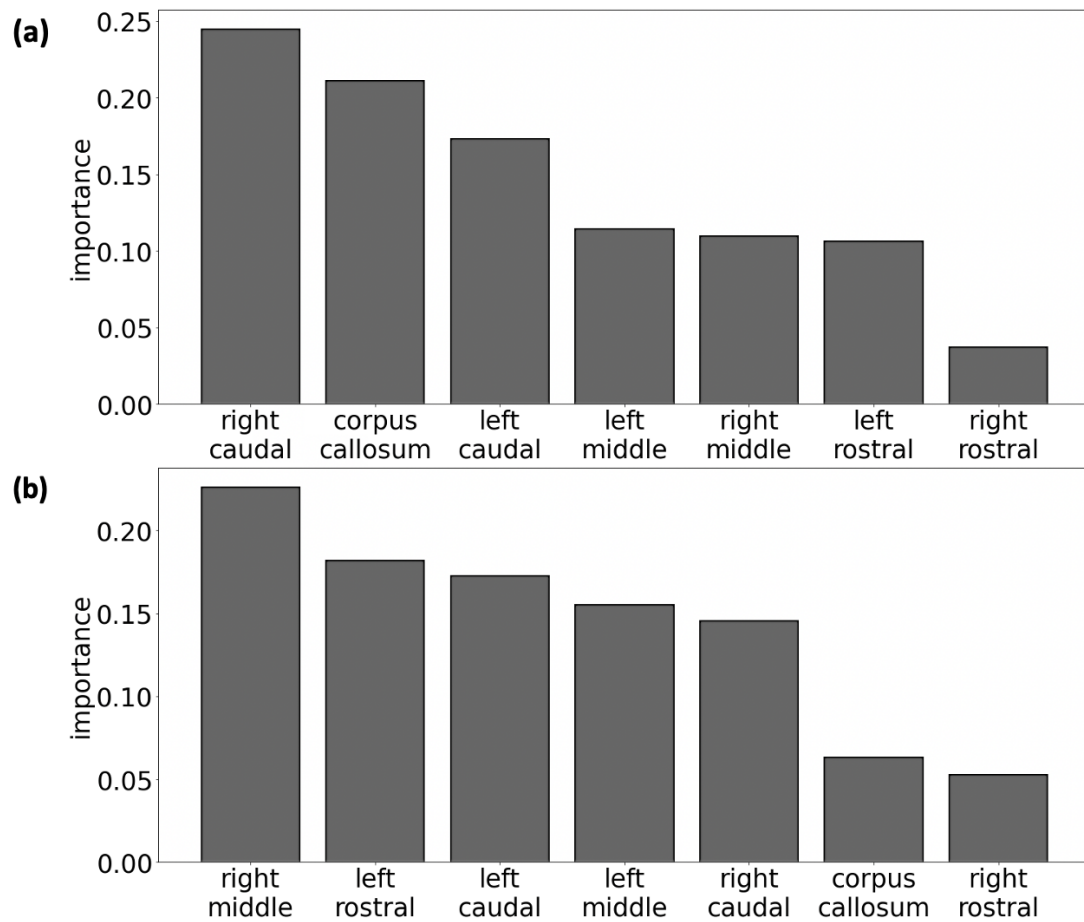


Figure 2.4: Relative importance of ROIs in identifying RHI-related diffusion changes; derived from the sum of the importance scores for each ROI across all diffusion metrics and all statistics: (a) logistic regression; (b) linear SVM. The relative importance of the various ROIs is not consistent between the two classifiers.

in such a population [12, 21, 46, 86, 107] as well as an association between altered MK and an athlete’s cumulative impact exposure [30, 35]. In the current study, WMTI metrics are found to be moderately important by both top-performing classifiers. Among WMTI metrics, relative importance of $D_{e,\perp}$ was consistently noted to be the highest, suggesting characteristics of the extracellular compartment as well as myelination are important rather than specific effects on neuronal axons themselves [66]. This seems reasonable, given that RHI are not associated with measurable, immediate focal neurological deficits though clearly requires further study.

Of note, for the top 2 classifiers, FA was not found to be especially informative despite prior reports associating FA changes with RHI [7, 12, 21, 30, 35, 46, 87, 107], (Figure 2). This may relate to known time-dependence of FA changes after injuries [124]: FA has been documented to change from elevated in the acute stage after MTBI to decreased in more chronic stages. The low importance of FA here may also be an artifact of dependency between FA and information already reflected in some of the other metrics.

In terms of brain regions and their relative importance in RHI, it appears that a regional approach leads to substantially better classification than merely using statistics computed over the whole brain WM. Despite the rather gross delineation of ROIs, we show that a region-based analysis adds value. This may be because localized changes are more difficult to detect once averaged across the entire WM. We also see that unlike the diffusion metrics, the relative importances of the 7 ROIs lack clear consistency between classifiers. In fact, 6 of the 7 ROIs show moderate importance in at least one of the top 2 performing classifiers. This may be due to the inherent heterogeneity of RHI in terms of location [30]. All 7 ROIs show

similar importance for the classification task, suggesting that the effects of RHI may be heterogeneous across individuals. Our finding is consistent with prior studies where various different brain regions have independently been associated with WM alterations relating to RHI [21, 30, 107].

The classifiers tested here were selected because they represent a range of types of classification algorithms as well as the span of learning capacity. We find that the two simple, linear classifiers (logistic regression and linear SVM) yield the best results, outperforming all nonlinear classifiers in both test AUC and cross-validation AUC by significant margins. The two trained linear classifiers also showed much smaller gaps between test and validation AUC, indicating good generalizability. If care is taken in the selection of classifiers, applying classic classifiers to medical imaging problems with limited but high-dimensional training data, good generalizability with robustness against overfitting can be achieved. Non-linear classifiers have the benefit of higher learning capacity but fail in this particular task due to poor generalizability on unseen data.

Using the wrapper-based feature selection method described and this representative group of complementary classifiers, despite the limitations of a small dataset and a high-dimensional feature space, our pipeline achieves AUC of 0.81 on a held-out test set. Pushing incrementally higher classification performance is not a focus of this work because the classification task is merely used here as a conduit to identify and understand important features of brain microstructure that may characterize exposure to RHI. We do, nevertheless, infer from an AUC of 0.81 that, in fact, there do exist some microstructural WM differences between contact sport athletes exposed to RHI and non-contact sport control athletes and that these differences can be detected using a simple linear classifier. This is in keeping with

a previous report examining statistical differences between these populations [30].

Limitations of this study include the small dataset in terms of number of subjects; however, we show proof of concept that the explicit feature selection through cross-validation approach used here can be effective and add to our understanding of underlying patterns even in a small medical imaging dataset. This may be a useful approach to study other practical questions in clinical cohorts. We find that simple linear classifiers are more appropriate given such a dataset, alleviating the potential of overfitting while still able to learn discriminative features [74]. Other limitations include heterogeneity of study athletes from a range of sports which are clearly not entirely equivalent in RHI exposure and RHI exposure was not directly quantified in these participants. Future work is already underway in the scientific community to better quantify RHI using helmet accelerators as well as machine learning computer vision models to analyze game videos [45, 101]. Future work could divide sports to subgroups based on quantitative RHI measurement once the RHI measurement and more subjects per sport become available. In addition, this study explicitly does not address RHI in female athletes as there were an insufficient number of appropriate female participants to account for sex-related differences. Female athletes are an important group to study as the effects of RHI may differ. Here, we did not use an exhaustive list of diffusion MRI metrics. The study serves as a pilot to show feasibility of using a classifier / feature-selection pipeline to better understand multidimensional and complex diffusion MRI. For the sake of this pilot, the diffusion model selected is a relatively simple yet established one that incorporates only a few general assumptions [111]. Finally, regional ROIs were based on a somewhat blunt division of 7 major WM regions so as to reduce dimensionality of the task. This may result in reducing the spatial specificity of the

diffusion measures. Future work could include smaller regions given larger datasets; nevertheless, this study presents the benefit of regional classifiers which show better performance than those using whole brain WM metrics.

2.2.5 Conclusion and Contributions

Our approach utilizes a classification and feature selection pipeline to unveil WM microstructural characteristics of RHI and identify important diffusion metrics. In this study, we have found measures of mean diffusivity, brain tissue microstructural complexity, and radial extra-axonal diffusion (MD, MK, and $D_{e,\perp}$) to be the three most relevant metrics that characterize subconcussive RHI. These pilot results do support the notion that there are detectable WM microstructure changes in the setting of subconcussive RHI exposure related to collegiate-level contact sport participation and that such changes may affect the extracellular space specifically. The type of approach taken here may be useful to better understand multidimensional, complicated diffusion MRI and the pathophysiologic ramifications in injury and disease.

Chapter 3

Deep Learning with Diffusion MRI as in vivo Microscope Reveals Sex-related Differences in Human White Matter Microstructure

3.1 Introduction

Biological sex (throughout this manuscript, sex, male and female refer to biological sex assigned at birth) is a crucial variable in neuroscience studies, for example, the National Institute of Health began to require reports of differences between males and females in all preclinical trials in 2014 [32]. Sex differences have been documented across a variety of cognitive functions such as motor cognitive performance [38, 88, 105], nonverbal reasoning [105], verbal working memory

Junbo Chen is the main driver of this study. Acknowledgment to Vara Lakshmi Bayanagari, Prof. Sohae Chung, Prof. Yao Wang, Prof. Yvonne W. Lui for their collaboration and advice.

[40, 68, 120], and episodic memory [5, 6, 60]. The prevalence of many neurological and neuropsychiatric disorders also differs between males and females: Autism spectrum disorder and Tourette syndrome are more prevalent in males [9, 123], while disorders such as multiple sclerosis and depression are more prevalent in females [97, 98]. Understanding sex differences in brain structure is crucial towards better understanding of sex differences in brain function and neuropsychiatric disorders. Prior studies of structural MRI have documented significant sex differences in global brain anatomy, such as greater overall brain volume in males compared with females [103]. Subregional brain differences have also been shown: males and females show differences in gray-matter volume across different brain regions [81], and investigators have also demonstrated greater cortical thickness in female subjects than males [102].

However, there is ongoing debate regarding sex-related differences in human brains, as the findings are not entirely consistent across studies. For example, there is controversy about sex differences in the size of the corpus callosum [1, 65, 82]. Some of these inconsistencies may be explained by variable quantification procedures, small sample sizes, and wide age distributions across studies [15].

Besides, the previous studies based on structural MRI focus on the sex differences in macroscopic brain structure. However, sex matters not only at the macroscopic level but also at the microscopic level [99]. Compared with macroscopic differences that inform gross brain structure, cellular-level changes at the microscopic level provide more information and more subtle indicators of cognitive function in both health and disease [10, 24, 42]. For example, cellular-level sex differences in the brain, such as the density of microglia, are critical for brain health and immunity and could influence differences in the sex-related expression of disease [50, 51].

Many such studies, however, rely primarily on animal models and the study of ex vivo samples which introduces fixation and preparation artifacts. As such, we have only a partial picture of human brain tissue microstructure. Elucidating sex-related brain tissue microstructural differences may help us better understand sex-related differences in normal development and aging as well as in pathologic conditions of the brain.

Multi-shell diffusion MRI is a promising and developing field capable of capturing microscopic structure of the brain non-invasively [94]. It is being used to study various neurological diseases, ranging from neurodegenerative disorders such as Alzheimer’s dementia [54, 130] and Parkinson’s disease [14] to autoimmune disorders such as Multiple Sclerosis [36]. As neurological disease disorders have been commonly documented with differences between males and females [9, 97, 98, 123], leveraging diffusion MRI to study the sex differences in the human brain could shed light on the sex-related differences on cellular level and help us understand pathology of neurological diseases.

In studies of differences between males and females in terms of diffusion MRI metrics, conventional statistical analysis methods are commonly used, such as comparing mean and variance of metric values within region-of-interest between sexes at group level based on t-test and f-test [67, 102], or comparing between males and females with tract-based spatial statistics analysis [109]. However, representing the brain microstructure with a few regional statistics is likely to lose complex and subtle microscopic information. Deep neural networks, on the other hand, are able to learn to capture complex biomarkers and non-linearity from the diffusion MRI volumes. However, training neural networks on high dimensional volumetric imaging data is challenging given limited dataset. Two recent works from two

different groups trained neural networks with handcrafted features derived from structural connectivity matrices computed using WM FA and mean diffusivity (MD) volumes. Their studies achieved between 77% - 95% accuracy in sex classification [56, 128], which suggests that there are indeed sex-related differences in structural connectivity. However, the use of complex hand-crafted features is cumbersome, can add potential biases, and limits the ease of reproducibility. Besides, different neural networks architectures are effective at capturing different types of features, making studies relying on one single architecture challenging to capture comprehensive information.

In our work, we aim to study sex differences in the human brain at microscopic level by leveraging deep neural networks and multi-shell diffusion MRI. We incorporate an end-to-end design wherein the neural networks take the entire MRI volume into account so as not to rely on complex hand-crafted feature engineering that bias analysis. In addition, we explore 3 major network architectures believed to capture different and possibly complementary information, to prevent the results from relying on a single model. Finally, we identify those WM areas that most significantly contribute to the model decisions, and thereby have most sex-related differences. With diffusion metrics of all subjects registered to a standard template space, effects of any macroscopic anatomical differences are removed to make the model focus on microscopic structural differences between sexes. Note that our effort in developing sex classification models is not for the purpose of classifying sexes per se, but rather to reveal how much in vivo signal from diffusion MRI may inform regarding sex-related differences in brain tissue microstructure.

3.2 Methods

3.2.1 End-to-End Classification Models

We test three major model architectures: 2D convolutional neural network (CNN), 3D CNN, and 3D vision transformer (ViT). We choose these end-to-end deep networks that act on the entire image volume to avoid any reliance on hand-crafted features and/or complicated feature engineering. In general, CNN and ViT show state-of-the-art performances broadly across image classification tasks and the two architectures have their own strengths and may be complementary: CNN has inductive bias by design such as locality and translation equivalence/invariance (w/wo pooling), making such a model generally more sample-efficient and easier in theory to capture local features of an image or volume [76]. While ViT lack the inductive bias from convolutional layers rendering them somewhat more data-hungry, they have strengths that CNNs lack in being able to capture long-range interactions and more global features present in an image or imaging volume [37, 39], which could be important [62, 104]. Thus, both CNN and ViT are included here. For the CNN, although 3D CNN may be an intuitive choice of architecture to handle a 3D imaging volume, a 3D CNN has many more parameters and requires more training samples compared with a 2D CNN. Therefore, we also assess the performance of a 2D CNN with a lighter feature extraction backbone and greater training efficiency.

3.2.1.1 2D Convolutional Neural Network

In this work, the 2D CNN employed uses a ResNet18 [58] as a backbone for feature extraction. Here, the 2D network essentially receives input from a small

3-slice subvolume (as ResNet18 is designed to receive color images with 3 channels (RGB)). Thus, we extract features from every 3 consecutive slices and combine features from all non-overlapping 3-slice subvolumes for the prediction head for classification (Figure 3.1). Specifically, given input volumetric data with the shape of $S \times H \times W$ (S: slice number, $H \times W$: slice size, with each slice in sagittal view), we generate $S/3$ 2D 3-channel images each with the shape of $3 \times H \times W$. The same ResNet18 is applied to extract features from each 3-slice subvolume and features from all $S/3$ 3-slice subvolumes are concatenated as the input to a linear prediction head. The ResNet18 architecture is shown at the bottom of Figure 1. The input is fed to a convolutional layer (conv layer) (kernel-size= 7×7 , stride=2, channel-number or number-of-feature-maps=64), followed by a max-pooling layer for further downsampling (kernel-size= 3×3 , stride=2). After the pooling, 8 convolutional layer blocks called residual blocks (where input to the block is added to the output via residual short-cut connection) are applied where each block contains 2 convolutional layers with kernel-size= 3×3 , the channel number gets doubled and the spatial size gets downsampled by 2 at the first conv layers of 3rd, 5th, 7th residual blocks. Each conv layer is followed by batch-normalization [63] and ReLU activation ???. In the end of ResNet18, global-average pooling is applied to each feature map to generate a single feature value, leading to 512 features for each 3-slice subvolume. Given $S \times H \times W = 183 \times 224 \times 224$, we have $S/3 = 61$ 3-slice subvolumes with each yielding 512 features. These 61×512 features are concatenated and fed to a linear layer for final prediction.

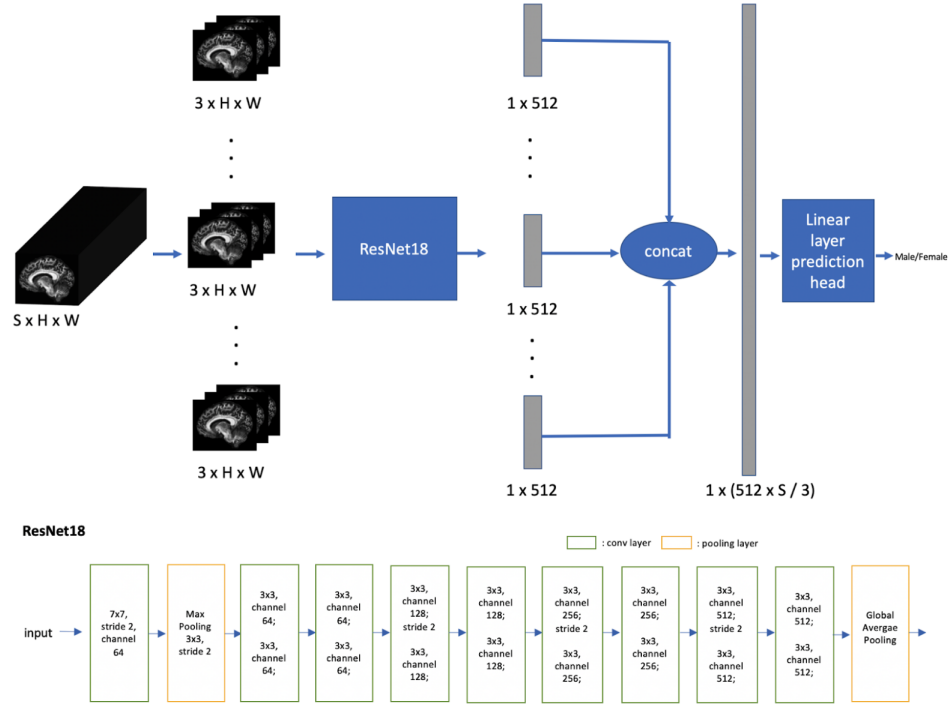


Figure 3.1: Our 2D CNN model. In the top of the figure, the imaging volume is divided into subvolumes, and a shared ResNet18 is applied to extract 512 features from each subvolume. The features are concatenated and fed to a linear layer for the final prediction. The bottom of the figure shows the architecture of ResNet18 (residual connection, ReLU activation, batch normalization are omitted for simplicity): The input is first fed into a convolutional layer (7x7 kernel-size, stride=2, channel-number=64) followed by a max-pooling (kernel-size=3x3, stride=2) layer; subsequently, 8 residual blocks are applied with each containing 2 convolutional layers. Residual blocks parameters: conv layers in block 1, 2 have kernel-size=3x3 and channel=64; conv layers in block 3, 4 have kernel-size=3x3 and channel=128; conv layers in block 5, 6 have kernel-size=3x3 and channel=256; conv layers in block 7, 8 have kernel-size=3x3 and channel=512; stride=2 is applied at the first conv layer of block 3, 5, 7. Global average pooling is applied at the end.

3.2.1.2 3D Convolutional Neural Network

We employ 3D ResNet-10 [22, 53] as our 3D CNN backbone, with architecture shown in Figure 3.2. The 3D volume is firstly fed into a conv layer (kernel-size=7x7x7, stride=2, channel=64) followed by a max pooling layer (kernel-size=3x3x3, stride=2), 8 residual blocks are then used with each block having 1 conv layer. The channel number is doubled at residual block 3, 5, 7, with stride set as 2 for block 3 and dilation set as 2 for block 5 and set as 4 for block 7. Each conv layer is followed by group-normalization [125] and ReLU activation [91]. In the end, global average pooling is applied to map 512 feature maps to 512 feature values and one linear layer is used for the final prediction.

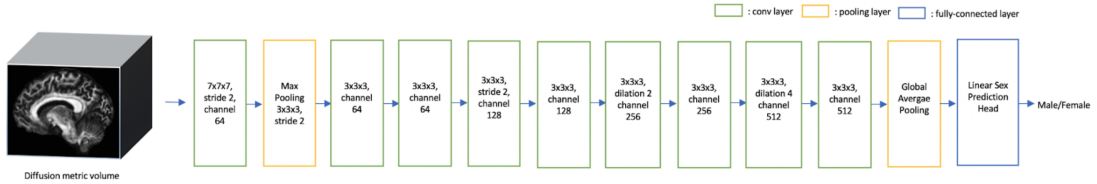


Figure 3.2: Our 3D CNN model based on ResNet10 (residual connection, ReLU activation, group normalization omitted for simplicity). The 3D volume is first fed to a conv layer (kernel-size=7x7x7, stride=2, channel=64) followed by a max pooling (kernel-size=3x3x3, stride=2). Subsequently, 8 residual blocks are applied with each containing 1 conv layer. Residual blocks parameters: block 1, 2 have kernel-size=3x3x3 and channel=64; block 3, 4 have kernel-size=3x3x3 and channel=128; block 5, 6 have kernel-size=3x3x3 and channel=256; block 7, 8 have kernel-size=3x3x3 and channel=512; stride=2 is used at conv layer in block 3, while dilation=2 is used at conv layer in block 5 and dilation=4 is used at conv layer in block 7. Global average pooling is applied at the end.

3.2.1.3 Vision Transformer for 3D Input Pretrained with Mask Autoencoders

The original 2D ViT [39] is extended to extract features from a 3D volume. Shown in Figure 3.3, given input 3D diffusion metric $x \in R^{S \times H \times W}$, the data is reshaped into a sequence of flattened non-overlapping 3D patches $x_p \in RN(shw)$, where (S, H, W) is 3D volume size and (s, h, w) is the 3D patch size, patch number is defined as $N = SHW/shw$. As shown in Figure 3, for each 3D patch, a linear layer is applied to map voxel values to a latent embedding with dimension D. A learnable positional embedding with same dimension D representing each token’s location, is added to the original embedding. The resulting sequence of embeddings for all N patches are fed to the encoder consisting of L alternating layers of multi-head attention and Multi-layer-perceptron (MLP) blocks. A classification token with dimension D is appended to the input embedding sequence, which is designed as a latent representing the entire input sample. The output embedding of the classification token is then fed into a linear prediction head to generate a prediction. In our study, $S \times H \times W = 182 \times 224 \times 224$ and $s \times h \times w = 6 \times 16 \times 16$, $D = 384$, $L = 12$.

We pretrain the ViT with a 2D+3D Masked Autoencoders (MAE) modified from 2D MAE [56], where a specific ratio of patches, defined as r , is randomly masked and a ViT encoder and auxiliary decoder are trained to predict the values of $r \cdot N$ masked patches from $(1 - r) \cdot N$ unmasked patches. After pretraining, the encoder is finetuned for the target sex classification task with all N patches fed into it. Since 3D patches are more difficult to predict than 2D patches (especially given the small number of available 3D volumes), we pretrain a 2D ViT encoder with MAE on 2D slices first and use the resulting weights to initialize our 3D ViT

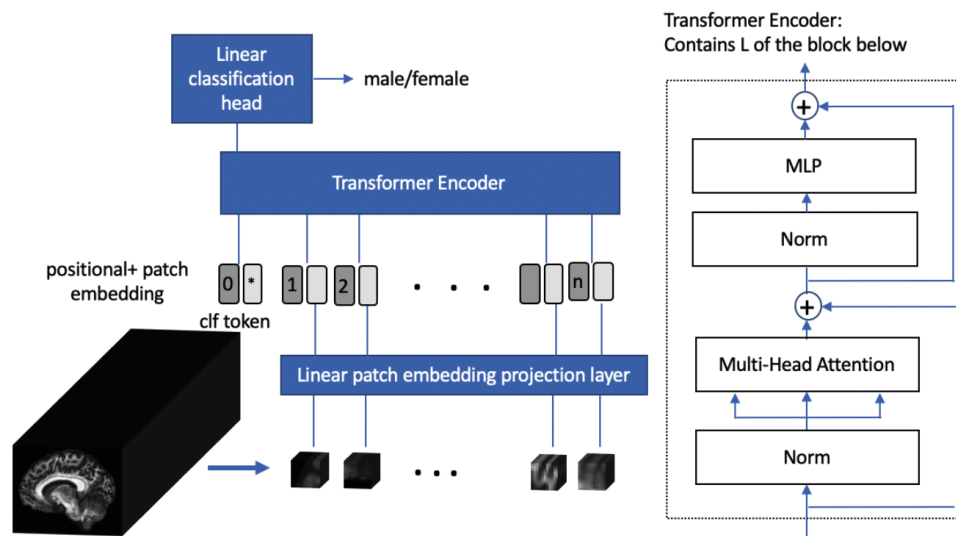


Figure 3.3: Vision Transformer for Diffusion MRI sex classification: the imaging volume inputted is partitioned into non-overlapping patches. Each patch is projected to patch embedding using a linear patch embedding layer, and added with positional embedding representing the position of the patch. A classification token is appended to the sequence of tokens to learn representation of the entire input sample. The structure of the transformer encoder is shown on the right, which consists of L alternating layers of multi-head attention and multiple-linear-perceptron (MLP) blocks. After the transformer encoder, the corresponding output of the classification token is fed to the classification head to generate prediction results.

model for 3D patches, and further pretrain the model with MAE on 3D volumes. In our study, mask ratio $r = 0.75$ and the axillary decoder has $D=192$, $L=4$. The ViT encoder (latent-dimension=384, depth=12, number of heads=6, MLP-ratio=4) and decoder (latent-dimension=192, depth=4, number of heads=3, MLP-ratio=4) adapt the asymmetric architecture following the MAE designed for images [56], as encoder only operates on visible patches and decoder operates on all patches during the pretraining, decoder is more memory consuming and should employ a smaller architecture. The masking ratio r is set as 75% as it was shown to be the best ratio for the image data, making the MAE task both feasible and challenging enough to learn generalizable features [56].

3.2.2 Model Training and Evaluation

1031 unique subjects are split into training (831 subjects), validation (100 subjects) and test sets (100 subjects). Training, validation and test sets share the same sex and age distribution, where female and male have a relatively balanced ratio of 27:23. Models' hyperparameters are tuned based on the performance on the validation set. Models trained with the training set and the selected hyperparameters are then tested on the test set for final prediction results. Classifiers are implemented with pytorch. For fair comparison, all three models use the same training/validation/testing split. Details of the training process are explained in the appendix. For ViT, we conducted three experiments: ViT trained from scratch without MAE pretraining, linear probing where the encoder is frozen with weights from MAE pretraining, and only linear prediction head is trained for sex classification, and fine-tuned ViT where the whole model is refined on sex labels. The performance of linear probing can reflect how the feature learnt from

pretraining generalizes to the sex classification task, while performance of model trained from scratch can serve as the baseline to examine if the pretraining can bring performance improvement.

For MAE, the model is trained with the mean square error between predicted pixels/voxels and their reconstruction target ground truth value for all masked patches. Instead of using the patches' original value, the reconstruction target is set as the values after z-score normalization with the mean and standard deviation of pixels/voxels in the patch, as this normalized target can help improve the representation quality of the pretraining [56]. The AdamW optimizer is used with $\beta_1=0.9$ and $\beta_2=0.95$. The weight decay is set as 0.05. For the 2D MAE, the model is trained for 500 epochs with batch size as 128 and initial learning rate as 7.5×10^{-5} . For the 3D MAE, the model is trained for 600 epochs with batch size as 8 and initial learning rate as 4.5×10^{-6} . For linear probing, logistic regression models from Scikit-learn [95] are trained on latents from frozen pretrained ViT encoder. For end-to-end finetuning, the model is trained for 100 epochs with initial learning rate as 1×10^{-4} and batch size 2 with cross-entropy as the loss function. For 3D CNN, the stochastic gradient descent optimizer is used with momentum as 0.9 and weight decay as 0.001. Exponential learning rate scheduler with $\gamma = 0.99$ is used. The model is trained for 100 epochs with initial learning rate as 0.01 and batch size 8. For 2D CNN, Adam optimizer is applied with learning rate at 0.03, momentum as 0.9 and beta values $\beta_1=0.9$ and $\beta_2=0.999$. The model was trained for 50 epochs with batch size set as 10.

3.2.3 Occlusion Analysis

We conduct occlusion analysis on the trained models and Wilcoxon signed rank test to identify white matter areas of the brain that contribute significantly to sex classification. We conduct occlusion at the region level and consider the 48 white matter regions defined by the Johns Hopkins University-ICBM-labels-1mm atlas [89]. Given a trained model for a diffusion metric, we compare the predicted probability for the correct label before and after occlusion of each region in succession, by setting all voxels in the region to the mean white matter value. We apply the Wilcoxon signed rank test with one-sided alternative hypothesis to the probability changes associated with each region for all subjects in the testing dataset to test whether the decrease in the predicted probability for the correct label is statistically significant. The regions that achieve p -value < 0.05 are considered significant for distinguishing between male and female.

3.3 Result

3.3.1 Diffusion MRI Acquisition and Processing

The study includes 1031 healthy adult subjects (age range, 22-37 years) from the Human Connectome Project (HCP - Young dataset) [117], whereby sex labels were collected through self-reporting and no subject was found to have different self-reported sex from genetic sex. Institutional review board approval and participants' informed consent were obtained at the participating institutions. Demographic details are summarized in Table 3.1.

Diffusion MR images were collected on a 3T scanner (Connectome Skyra,

Table 3.1: Study Cohort

	Male	Female
Number of subjects	471	560
Age range: number of subjects		
22 - 25	138	79
26 - 30	205	249
31 - 37	128	232

Siemens Medical Solutions, Erlangen, Germany) and preprocessed as per HCP protocol [48, 117]. In brief, diffusion imaging was performed with the following parameters: 3 b-values (1000, 2000, 3000 s/mm²), 90 diffusion orientations per shell, 18 b₀ (b-value=0) images, 1.25 mm isotropic image resolution, field of view = 210 mm, number of slices=111, TR/TE=5500/89 ms, each scan was repeated along 2 phase encoding directions (RL/LR), details can be found in HCP dataset [117]. The diffusion data was preprocessed by HCP for correction of artifacts like motion and eddy-currents artifacts, detailed in [48]. We use a state-of-the-art image processing pipeline to generate diffusion metrics [2]. We use tissue diffusion anisotropy (FA, fractional anisotropy), mean diffusivity (MD) from Diffusion Tensor Imaging (DTI) and tissue complexity (MK, mean kurtosis) from Diffusion Kurtosis Imaging (DKI) to assess white matter microstructure. FA and MD are included because they are the two most commonly used metrics for characterization of tissue microstructure in brain-related studies [119]. Of note, FA measures directionality of water movement in brain tissue, known to be sensitive to microstructures such as axons and myelin [113]; and MD measures mean water diffusivity, sensitive to characteristics like cellularity [93]. Here, we also include MK from DKI to compactly represent non-Gaussian water diffusivity as a measure of overall tissue

microstructural complexity [30]. All metrics are registered to standard MNI space [89] using FMRIB Software Library (FSL) [111] so as to remove effects of any macroscopic anatomical differences such as size and contour of the brain itself.

3.3.2 Classification Results

We use the area under the curve (AUC) of each trained model on the testing dataset to evaluate the model performance. Table 3.3 shows that our 2D CNN, 3D CNN and ViT (finetuned and linear probing) models all achieved promising AUC for all 3 diffusion metrics with test AUC of >0.9 . For FA and MD, 2D CNN achieved the highest AUC at 0.98 for FA and at 0.97 for MD. 3D CNN and ViT also achieved relatively high AUC (> 0.92). For MK, all models achieved a high AUC above 0.96, and 3D CNN achieved highest performance with AUC of 0.98. The ViT trained from scratch yielded low AUC (< 0.8) for all diffusion metrics. The finetuned ViT and linear probing ViT achieved comparable AUC on all 3 diffusion metrics, indicating that the MAE-pretrained feature extraction layer is directly applicable for the classification task.

Table 3.2: Performance (test AUC) of sex classification models using three different diffusion MRI parametric maps as inputs (FA, MD, and MK)

Model	FA	MD	MK
2DCNN	0.98	0.97	0.96
3DCNN	0.92	0.96	0.98
ViT (finetuned)	0.93	0.95	0.97
ViT (linear probing)	0.94	0.95	0.96
ViT (trained from scratch)	0.79	0.75	0.72

3.3.3 Occlusion Analysis Results

2D and 3D CNNs and finetuned ViT are included in the occlusion analysis. The ViT finetuned model is selected for the occlusion analysis despite it has similar performance as the linear probing model, because the finetuned model is refined on the sex classification task. The numbers of regions passing the significance test are summarized in Table 3.3. Identified regions are illustrated in Figures 3.4-3.6.

Table 3.3: Number of white matter regions showing significant differences between males and females in the occlusion analysis; 48 WM regions in total.

Model	FA	MD	MK
2DCNN	12	25	7
3DCNN	2	2	2
ViT (finetuned)	5	13	2

3.4 Discussion

The study reveals clear sex-related differences in white matter microstructure as captured by diffusion MRI, detected consistently across 3 different end-to-end, deep learning-based image classification models. The reliability of this finding is evident in the fact that high classification performance (test AUC 0.92 - 0.98) is observed independent of model architecture across 3 major network architecture types and without introducing the biases of complex hand-crafted features and/or manual operations. In addition, white matter regions most central to the model decision are identified and may shed some additional light on these differences.

The three different model architectures allow us to leverage different types of features for sex classification. For example, the 3D CNN relies on a conventional 3D

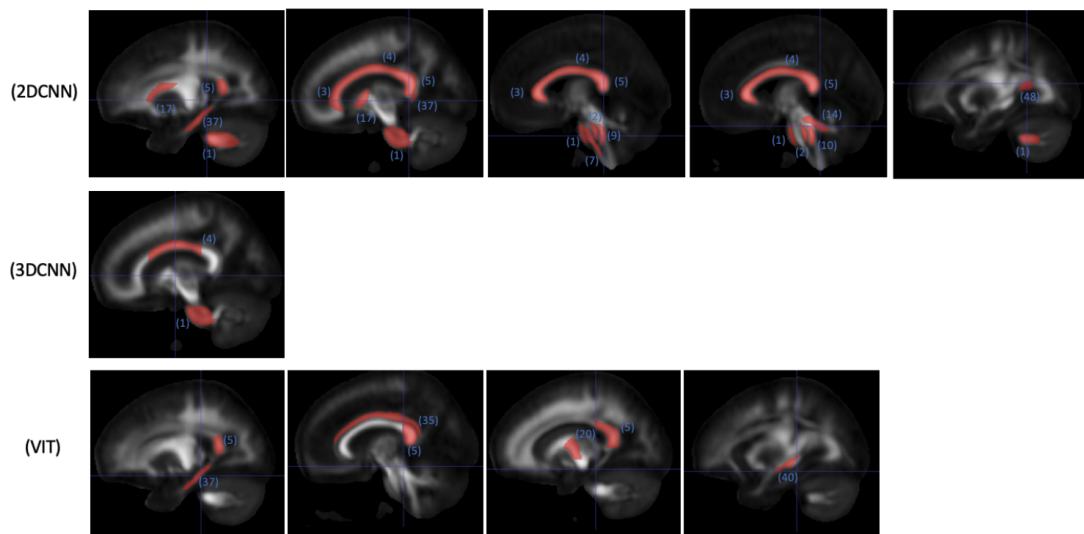


Figure 3.4: WM regions in selected slices with significant ($p < 0.05$) impact on classification probability based on occlusion analysis for FA; numbered labels based on JHU-ICBM-1mm atlas (<https://identifiers.org/neurovault.image:1401>); 1: middle cerebellar peduncle, 2: pontine crossing tract (a part of middle cerebellar peduncle), 3: genu of corpus callosum, 4: body of corpus callosum, 5: splenium of corpus callosum, 7: corticospinal tract (right), 9: medial lemniscus (right), 10: medial lemniscus (left), 14: superior cerebellar peduncle (left); 17: anterior limb of internal capsule (right), 20: posterior limb of internal capsule (left), 35: cingulum (cingulate gyrus) (right), 37: cingulum (hippocampus) (right), 40: stria terminalis (left), 48: tapetum (left).

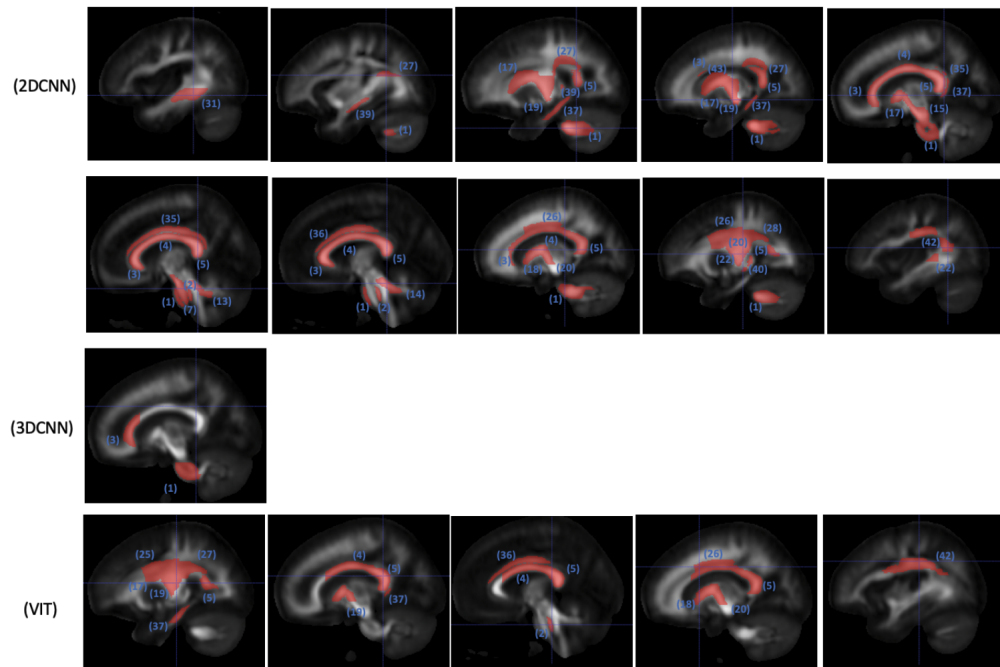


Figure 3.5: WM regions in selected slices with significant ($p < 0.05$) impact on classification probability based on occlusion analysis for MD; numbered labels based on JHU-ICBM-1mm atlas (<https://identifiers.org/neurovault.image:1401>); 1: middle cerebellar peduncle, 2: pontine crossing tract (a part of middle cerebellar peduncle), 3: genu of corpus callosum, 4: body of corpus callosum, 5: splenium of corpus callosum, 5: splenium of corpus callosum, 7: corticospinal tract (right), 13: superior cerebellar peduncle (right), 14: superior cerebellar peduncle (left), 15: cerebral peduncle (right), 17: anterior limb of internal capsule (right), 18: anterior limb of internal capsule (left), 19: posterior limb of internal capsule (right), 20: posterior limb of internal capsule (left), 22: retrolenticular part of internal capsule (left), 25: superior corona radiata (right), 26: superior corona radiata (left), 27: posterior corona radiata (right), 28: posterior corona radiata (left), 31: sagittal stratum (right), 35: Cingulum (cingulate gyrus) (right), 36: cingulum (cingulate gyrus) (left), 37: cingulum (hippocampus) (right), 39: stria terminalis (right), 40: Stria terminalis (left), 42: superior longitudinal fasciculus (left), 43: superior fronto-occipital fasciculus (right).

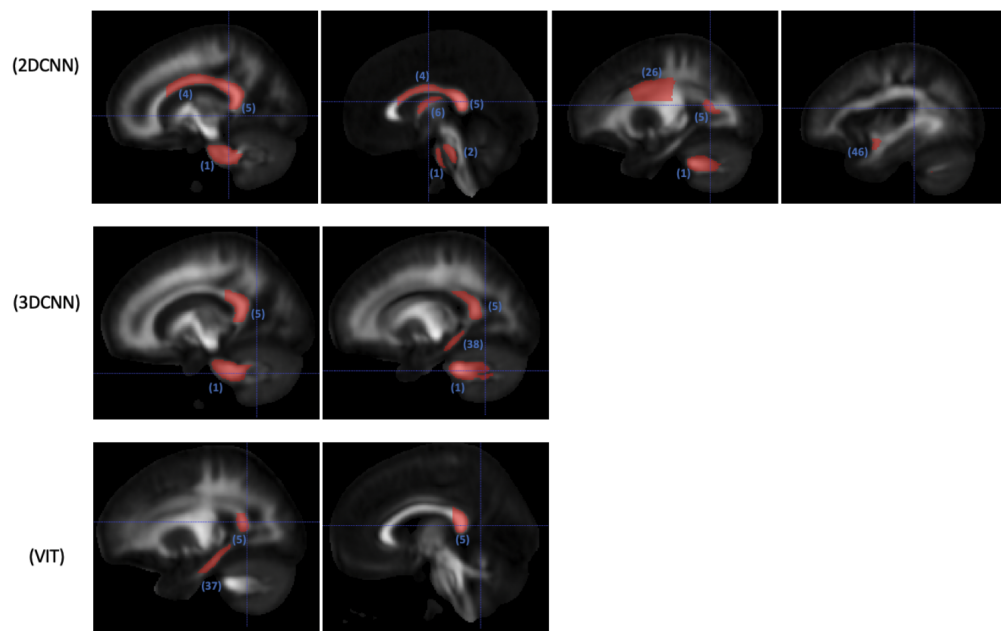


Figure 3.6: WM regions in selected slices with significant ($p < 0.05$) impact on classification probability based on occlusion analysis for MK; numbered labels based on JHU-ICBM-1mm atlas (<https://identifiers.org/neurovault.image:1401>); 1: middle cerebellar peduncle, 2: pontine crossing tract (a part of middle cerebellar peduncle), 4: body of corpus callosum, 5: splenium of corpus callosum, 6: fornix (column and body of fornix), 26: superior corona radiata (left), 37: Cingulum (hippocampus) (right), 38: cingulum (hippocampus) (left), 46: uncinat fasciculus (left).

CNN backbone and while it can powerfully capture local features within the imaging volume, a recent study showed that even very deep CNNs still have only small effective receptive fields ([37]), meaning they are better able to learn local features as opposed to longer distance relationships. On the other hand, the nature of the ViT enables it to capture global features more readily [39] and the incorporated MAE pretraining task used here also heavily focuses on inter-patch correlations. As both the 3D CNN and ViT models performed very well, this suggests that there are perhaps both short-distance and long-distance interactions contributing to sex-related differences in white matter microstructure.

Of note, the 2D CNN achieved overall best performance for 2 out of 3 diffusion metrics studied. This could be attributable to two main factors: First, the 2D CNN model used the simplest feature extraction backbone with the lowest number of parameters, possibly pushing its generalization capability given a somewhat modest-sized training dataset. However, the differences between validation and test performances were nominal for all three models, suggesting generalizability of the models to be comparable. One additional consideration could be that the 2D CNN classifier used here incorporates a design that may allow it to simultaneously capture both local features and global interactions (across all slices), thus making it able to leverage both types of features in the classification task. Specifically, the ResNet18 extracts features from every group of 3 consecutive slices allowing the model to learn from within-slice features and short-range inter-slice features across the 3 slices; by then concatenating features across all 3-slice subvolumes (as opposed to averaging across them as is most commonly done) the model here effectively preserves local features from every 3-slice partition while at the same time, the prediction head is able to learn more global interactions across 3-slice

subvolumes.

The occlusion analysis results show general consistency across models and across diffusion metrics and implicate central white matter tracts and ventral/dorsal hindbrain tracts in contributing to sex-related differences, though results differ slightly across the three diffusion metrics and the three models tested. Of interest, the number and fractional volume of WM regions significantly contributing to sex classification was highest for 2D CNN (mean number of regions: 15; mean fractional volume: 0.79) compared with 3D CNN (mean number of regions: 2; mean fractional volume: 0.24) and ViT (mean number of regions: 7; mean fractional volume: 0.16), possibly again reflecting differences in the relative facility of these models to tap short-range interactions, long-range interactions, or both. Across the three diffusion metrics, it appears that the 3D CNN classifier focused consistently on large central white matter structures such as the middle cerebellar peduncle and corpus callosum whereas the ViT and 2D CNN models tended to rely on a greater diversity of white matter regions. Another observation is that corpus callosum was found to be important across all three neural networks architectures and three included diffusion metrics. As there is debate whether sex differences exist within the corpus callosum [1, 65, 82], our work provides new evidence that sex differences exist in the corpus callosum.

For the ViT, pre-training with MAE was important. ViT is a data-hungry architecture and difficult to train with a limited dataset since it lacks inductive bias such as the locality and translation invariance of CNNs. The MAE pretraining task (to predict masked patches from visible patches) enables the model to learn inter-patch interactions without supervision from data labels. The random masking itself also introduces data diversity to the pretraining, which helps further improve

the generalizability of learned features. The benefit of MAE pretraining is clearly demonstrated in the experimental results: without pretraining, ViT trained from scratch yielded much lower performance with test AUC < 0.80 . With MAE pretraining, the ViT encoder achieved test AUC 0.94-0.96. The end-to-end supervised finetuning brought no additional gain and achieved comparable performance with linear probing, confirming that the size of the training set is insufficient to tune a data-hungry ViT in supervised end-to-end training.

Our results demonstrate that microstructural sex differences exist in the human brain both in local features (e.g., within central white matter structures such as the middle cerebellar peduncle and corpus callosum) and in global features (like long-distance interactions). Capturing microstructural differences with such complexity is very challenging for conventional statistical methods or a single neural network architecture. Our work shows that, instead of relying on a single neural network architecture, leveraging multiple neural networks with very different architecture design can capture complementary information and make the results independent of the model architecture. When it comes to leverage data-hungry neural network architectures for additional information, self-supervised learning can be used to pretrain the models and enable these neural networks applicable to medical imaging studies that lack large datasets. In summary, our work provides an example of a framework to study sex differences in the human brain at microscopic level with multiple deep neural networks and multi-shell diffusion MRI, which can capture complex features that reflect sex differences and prevent results from being biased by the model type. Such a framework can be further applied to study the brain microstructure underlying neurological diseases such as neurodegenerative disorders like Alzheimer’s dementia and Parkinson’s disease or autoimmune disorders such

as Multiple Sclerosis, which have been found to manifest differently between males and females [9, 97, 123].

Limitations include the use of only three representative diffusion metrics, though these were chosen based on the fact that they are common and easily obtained. Further exploring modeled diffusion metrics [94] may yield more information about sex-related differences in tissue microstructure and help us better characterize the underlying biophysical differences between brains of males and females. Recognizing that the age distribution differs between the female and male cohorts (with the female group having more older people) (3.1), we have separately evaluated the model accuracy on the three age groups and found the accuracies to be comparable among these groups. Even for the middle age group (26-30), our models achieve high sex classification accuracy, thus affirming that our models are not mostly using microstructure differences due to age to separate different sex groups. Finally, the occlusion analysis was conducted using a standard JHU-ICBM-1mm atlas for white matter parcellation with sizable variation in region size which could potentially bias regional importance; however, our analysis shows that the significance of regions is not merely based on the region size.

3.5 Conclusion and Contributions

This study finds that there are clear sex-related differences in the brain white matter microstructure of healthy young adults that can be detected in vivo using diffusion MRI without hand-crafting or manually manipulating the imaging data. We show this utilizing 3 different end-to-end deep neural networks and 3 diffusion MRI metrics. Even after registering diffusion MR volumes to a template so as to

remove macroscopic anatomical differences such as overall brain size and contour, we find that sex differences exist in diffusion anisotropy (FA), mean diffusivity (MD) and tissue complexity (MK) of brain white matter. Our experiments further suggest that there are both local as well as longer-distance microstructural organizational features that differ between sexes. In particular, the central white matter appears specifically implicated. This study provides a framework to study microstructural differences in the human brain with multiple neural network architectures, which help capture complex microscopic features challenging for statistical methods while preventing the results from depending on a single model. Further study is needed to determine whether and how these microstructural differences influence brain health and disease in both men and women.

Chapter 4

Temporal Swin Transformer for Grid-Free ECoG Speech Decoding on Single and Multi Patient

4.1 Introduction

The speech disability can seriously decrease the patient’s life quality and can be caused by brain damage such as stroke, brain injury and tumor [29, 92, 115]. In the United States, an estimated 2.5 million people are suffering from the disability of speech [64]. The electrocorticographic (ECoG) signals can record the neural activity of speech production and be used to generate human speech, making it possible to design Brain-computer interface to help patients with speech disability to communicate [20, 23, 83, 90, 100, 108].

Junbo Chen is the main driver of this study. Acknowledgment to Xupeng Chen, Dr. Ran Wang, Dr. Amirhossein Khalilian-Gourtani, Chenqian Le, Prof. Adeen Flinker, and Prof. Yao Wang for their collaboration and advice.

The recent advancements in deep neural networks can be leveraged to push the boundary of speech decoding from ECoG signals. In [28, 122], the ResNet [58] and 3D Swin Transformer [79] were used as ECoG decoder to predict time-varying speech parameters, and achieved promising performances. In [3], densely connected 3D Convolutional Neural Networks (CNN) was applied to decode speech from ECoG signals. Besides CNN and Transformer, Recurrent neural networks (RNN) and long short term memory (LSTM) have also been explored as ECoG decoder [4, 73, 84]. These recent studies demonstrate that deep neural networks are capable of decoding speech information from the complex neural activity recorded by the ECoG signals.

However, the deep neural networks in previous ECoG studies have architecture designs that require the electrodes to be put in a fixed grid topology, which imposes a major challenge to fully leveraging the ECoG signals. Firstly, the ECoG electrodes are implanted in the human not necessarily following any fixed grid. For a single subject, the electrodes could be implanted with grids and strips in very different positions. Additionally, the electrodes can be implanted below the brain surface as depth electrodes. The neural networks such as CNN require electrodes to be fit in a fixed grid [3, 4, 73, 84, 122, 127], making the model not able to leverage electrodes that can not fit in the grid. Vision transformers’s absolute position embeddings and relative positional bias are also based on the grid index [39, 77, 78, 79, 122, 127]. Besides, the position of implanted electrodes has high variation among subjects. The deep neural networks that based on fixed grid topology can not handle the subject differences. Therefore, previous studies rely on the subject-specific model, making the ECoG decoder not able to leverage signals from multiple subjects and generalize to new subjects outside of the training set.

In our study, we propose a novel transformer-based ECoG decoder that does not rely on regular grid structure, named non-grid Swin transformer with temporal window (SwinT). Instead of relying on the grid index, the model leverages the anatomical location of electrodes in the standardized brain template to help the prediction of speech. The proposed ECoG decoder achieved superior performances than ResNet and 3D Swin Transformer, given the same electrodes. The model demonstrated further performance increase by leveraging the off-grid electrodes that can not be used in the previous studies. Besides, instead of relying on the subject-specific ECoG decoder, we managed to train a single model with multiple subjects, leading to performance improvement and generalizability to unseen subjects that are not in the training set.

4.2 Method

4.2.1 Speech Decoding Framework

The design of the ECoG-to-Speech framework is based on the 2-step speech decoding framework proposed by our previous study [28], shown in Figure 4.1. In the first step of Audio-to-Audio training (upper part of Figure 4.1), a speech encoder is used to extract speech parameters at every time frame (e.g. pitch, formant frequencies, loudness) from input speech spectrogram and a differentiable speech decoder/synthesizer is designed to reconstruct the spectrogram from the speech parameters. In the second step of ECoG-to-Audio training (lower part of Figure 4.1), the ECoG Decoder is trained to predict the time-varying speech parameters from ECoG signals. The speech parameters generated by the ECoG Decoder will be fed to the Speech Synthesizer from the Audio-to-Audio training to

generate a speech spectrogram that will be converted to the final predicted speech waveform.

Following [28, 122], for Audio-to-Audio training, in the speech synthesizer, the speech signals can be parameterized as a soft mix of voice contents and unvoice contents: voice content is generated by processing a harmonic excitation with a voice filter designed as the sum of 6 formant filters, designed to model formants such as vowels and nasal information; unvoice content is generated by processing white noise with a broadband filter as well as the six formant filters, designed to capture consonants (such as fricatives, plosives, semi-voice, and unvoice) and formant transition subsequent to consonants. The weighted average of the voice and unvoice contents is then modulated with loudness and added with background noise to generate the final speech spectrogram. Based on the design of the Speech Synthesizer from our previous study [28], the speech signal is parameterized as 18 time-varying speech parameters: fundamental frequency of harmonic excitation f_0^t , formant frequency f_i^t and amplitude a_i^t of each of the six formant filter, center frequency f_u^t , bandwidth b_u^t and amplitude a_u^t defining the broadband unvoice filter, voice weight α^t (α^t for voice component and $(1 - \alpha^t)$ for unvoice component) and loudness L^t . As shown in the upper part of Figure 4.1, during Audio-to-Audio training, the speech encoder extracts 18 speech parameters at each time from the original speech spectrogram, and the set of speech parameters is then fed to the Speech synthesizer to reconstruct the original speech spectrogram. A simple network architecture made with MLP and temporal convolution is applied for the speech encoder. The soft mix of voice and unvoice components makes the speech synthesizer differentiable, which enables the end-to-end training of this speech-to-speech autoencoding tasks. The speech encoder and synthesizer details

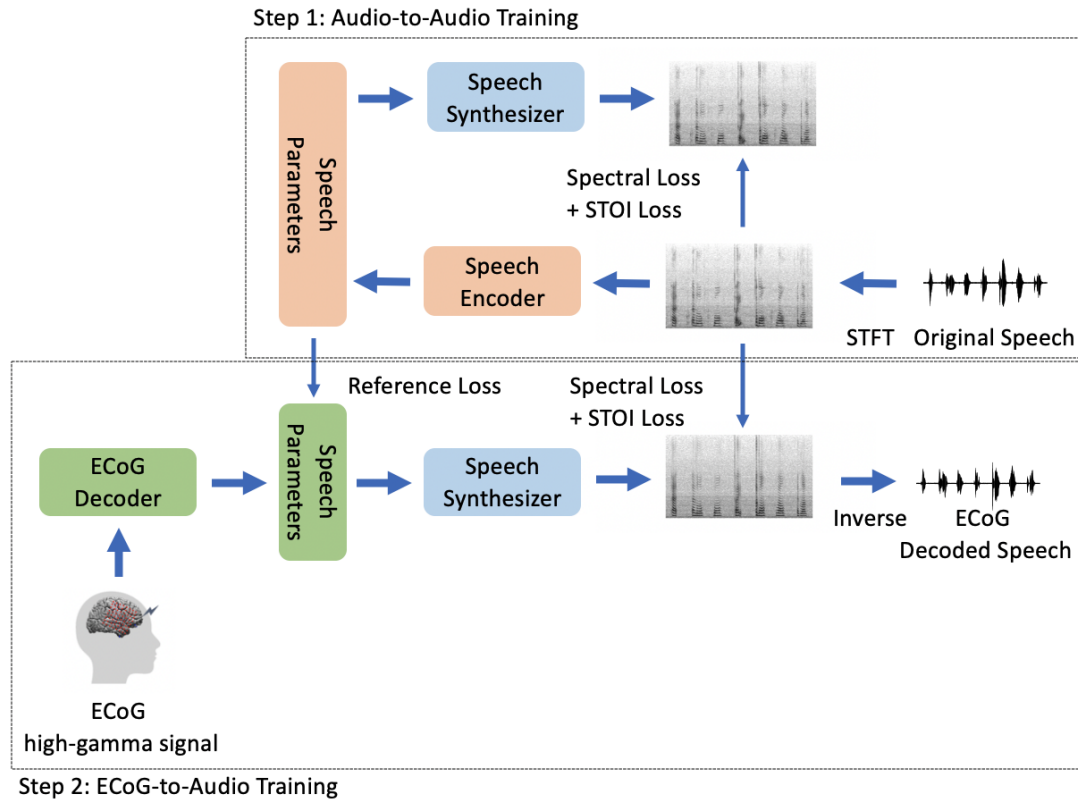


Figure 4.1: 2-Step Speech Decoding Training Framework (step1) Audio-to-Audio Training: the original speech waveform is first converted to speech spectrogram with STFT, then speech parameters are generated at each frame based on speech spectrogram by Speech Encoder. A speech synthesizer is trained to reconstruct the speech spectrogram based on speech parameters from the Speech encoder. (step2) ECoG-to-Audio Training: The ECoG decoder maps ECoG high-gamma signal to latent representation and predicts speech parameters supervised by the speech parameters from the trained speech encoder from step 1. The predicted speech parameters from the ECoG decoder are fed to the trained speech synthesizer from step 1 to generate the predicted speech spectrogram, which is reversed to the predicted speech signal.

can be found in [28].

For ECoG-to-Audio training, the ECoG decoder first maps neural activity from all input electrodes to a latent representation and predicts the 18 speech parameters for each time frame, supervised by the speech parameters generated by the Speech Encoder from Audio-to-Audio training. Then, the speech parameters predicted by the ECoG decoder will be fed into the Speech Synthesizer from the Audio-to-Audio training to generate the predicted spectrogram, which is converted back to the ECoG-decoded speech signal.

4.2.2 Grid-Free ECoG Decoder based on Temporal Swin Transformer

In our study, we propose a novel architecture for decoding speech parameters from ECoG signals that does not require electrodes to be formatted as a 2D grid. We name the proposed ECoG decoder as non-grid Swin transformer with temporal window (SwinT), which is inspired by the Swin Transformer [77, 78]. In the vanilla Vision Transformer (ViT) [39], the self-attention layer computes global attention among all tokens (with each token corresponding to an image patch). This global attention causes the absence of the inductive bias of locality and heavy quadratic computational complexity to the input size. The Swin Transformer solves the problems by partitioning tokens into local windows and computing local attention within each window at self-attention layers. To allow inter-window information exchange, the Swin Transformer shifts the window partition for every two windowed self-attention layers, which prevents different windows from being segregated (details can be found in [77, 78]). However, since the Swin Transformer was designed for 2D images or 3D videos, its architecture assumes the input is in the formats of 2D

or 3D grids. In our proposed SwinT, we made several modifications to remove such a constraint to allow speech decoding based on ECoG electrodes in any topological layout. The architecture of the SwinT is shown in Figure 4.2.

Grid-free patch partition: In the Swin Transformer [77, 78, 79] or ViT [39], the input images or videos are partitioned into 2D or 3D patches, and each patch is then mapped to a token with a patch embedding layer. This patch partition enforces the assumption of grid input and makes the model not invariant to the electrode order (as changing the order will change the electrodes corresponding to each patch). To solve this problem, our proposed SwinT generates tokens from each electrode individually and only partitions the temporal dimension. As shown in Figure 4.2, given an ECoG signal with the shape of $T \times N$ (T : number of frames, N : number of electrodes), for each electrode, the SwinT partitions the temporal sequence of neural activity as $\frac{T}{W}$ patches with patch size W . The temporal patch partition generates $\frac{T}{W} \times N$ patches in total, and a linear patch embedding layer is applied to map them to $\frac{T}{W} \times N$ tokens with latent dimension of C .

Temporal window attention: In Swin transformer [77, 78, 79], tokens are partitioned into windows, where each window contains a local subset of tokens, and attention is calculated among tokens within each window. The window partition and local attention in spatial dimension make the model only suitable for signals in grid format. In SwinT, to remove this grid input assumption, the model only partitions tokens into local windows in the temporal dimension and calculates global attention in the spatial dimension. Given $N = N_t \times N_s$ tokens (N : total number of tokens, N_t : length of tokens in the temporal dimension, N_s : length of tokens in the spatial dimension) and window size W_t , the N tokens are partitioned into $\frac{N_t}{W_t}$ windows and attention is calculated among $W_t \times N_s$ tokens within each

window.

Temporal patch merging: The Swin Transformer leverages patch merging to achieve inductive bias of locality and hierarchical feature maps. However, merging nearby patches in the spatial dimension does not fit our study as it enforces the grid input assumption. Therefore, instead of using the spatiotemporal patch merging in the 3D Swin Transformer [79], the SwinT conducts patch merging for each electrode individually. During patch merging in the SwinT, for each electrode, every two consecutive tokens in the temporal dimension with shape C will be concatenated as a $2C$ dimensional latent and get mapped to a $2C$ dimensional merged token.

Grid-free positional embedding: The SwinT follows Swin Transformers [77] to add positional information as relative positional bias. However, instead of using the 2D or 3D grid index difference as the relative position like the Swin Transformer, our SwinT defines the relative positional bias based on anatomical location and time-frame index of each token. The positional bias is defined as below:

$$Attention(Q, K, V) = Softmax(SIM(Q, K))V \quad (4.1)$$

$$SIM(q_i, k_j) = \frac{q_i k_j}{|q_i| |k_j|} / \tau + B_{i,j} \quad (4.2)$$

$$B_{i,j} = MLP(x_i, y_i, z_i, t_i, x_j, y_j, z_j, t_j, x_i - x_j, y_i - y_j, z_i - z_j, t_i - t_j) + r_i \cdot r_j \quad (4.3)$$

Given $Q, K, V \in R^{N \times C}$ (Q, K, V are query, key and value generated from each token, N is number of tokens and C is the latent dimension), shown in equation 4.1, the softmax of $SIM(Q, K)$ for all pairs of token in the window is used to aggregated V (values of tokens within the window) to get the output token values. We define query-key similarity following the scaled cosine attention of SwinV2 [77], defined in equation 4.2. τ is a learnable parameter not shared among attention

heads and layers. $B_{i,j}$ is the relative positional bias between token i and token j . In SwinT, $B_{i,j}$ consists of two terms: MNI-based positional bias and region-index-based bias. For MNI-based positional bias, we project each subject’s electrodes to a standardized Montreal Neurological Institute (MNI) brain anatomical map and collect each electrode’s x, y, z location in standard MNI coordinate. For each token pair, the MNI coordinates of the corresponding electrodes and time-frame index, along with the difference, will be mapped to the MNI-based positional bias with a 2-layer MLP, which is shown in the first term of equation 4.3. Besides, we parcellate the brains into region-of-interest (ROI) and learn a dictionary of embeddings for all ROIs. Given N_r ROIs and N_h attention head, the learnable dictionary has N_h sets of $N_r \times C_r$ region embeddings (C_r is the region embedding dimension). The region embeddings are learnt during the training. For a pair of tokens, the dot product of the embeddings of their corresponding electrodes’ ROIs will be added to the positional bias, shown in the second term of equation 4.3. The dot product is used instead of cosine similarity to remove inductive bias of strong intra-region attentions, as the inductive bias may not help the speech decoding. Besides, it can also allow the model to assign high attention to certain regions by letting them have large embedding values.

The architecture of SwinT is shown in Figure 4.2 (a). The input ECoG signal with a size of $T \times N$ is partitioned into $(\frac{T}{W} \times N)$ patches, each with a patch size of $W \times 1$. A linear patch embedding layer then maps each patch to a C dimensional token. The SwinT has three stages with 2, 2, and 6 layers, respectively. Swin Transformer Block (consists of a windowed multi-head self-attention layer and an MLP) is applied in each layer, detailed in [78], and we replace the spatial-temporal windowing with temporal-only windowing. Following the Swin Transformer, for

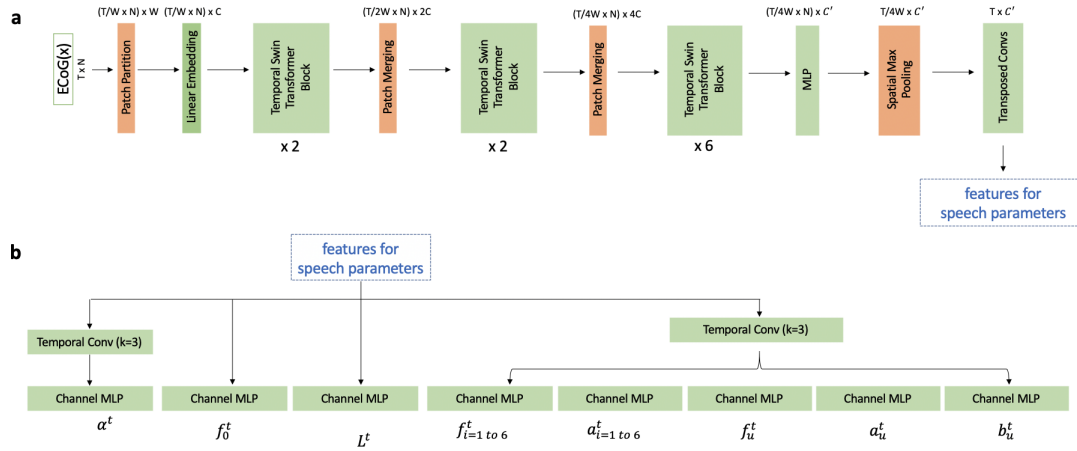


Figure 4.2: **a**. SwinT ECoG decoder. SwinT uses three stages of temporal swin transformer blocks with spatial-temporal attention with temporal windowing to extract features. An MLP layer is applied to decrease latent dimension. Spatial max pooling is then applied followed by transposed temporal convolution to upsample the temporal dimension. **b**. Prediction head for speech parameters consists of temporal convolution and MLP that map features from (a) to speech parameters at every frame.

every two consecutive layers, the second layer will shift the window partition to allow inter-window information exchange, detailed in [78]. SwinT performs temporal patch merging after the first and second stages, which decreases the token number by half and doubles the latent dimension. After stage 3, an MLP is applied to decrease $4C$ latent dimension to C' . Spatial max pooling is then applied to convert $(\frac{T}{4W} \times N) \times C'$ feature maps to $\frac{T}{4W} \times C'$, followed by transposed temporal convolutions to upsample $\frac{T}{4W} \times C'$ to $T \times C'$, where T is the frame number of input ECoG signal. As shown in 4.2 (b), the $T \times C'$ latent from SwinT goes through prediction head consists of temporal convolutions (kernel-size=3) and MLP, proposed in previous work [28], to predict the 18 speech parameters at every frame, which will be used to generate the speech spectrogram with the speech synthesizer from the audio-to-audio training.

In our study, we set $C = 96$ and $C' = 32$. Patch-size $W = 4$ is applied to partition temporal dimension. In our 3 stages SwinT with 2, 2, and 6 layers, the self-attention layers in the 3 stages have 3, 6, and 12 attention heads, respectively. The MLP ratio of transformer blocks is set as 4. The MLP for dimension decrease has 3 layers ($384 \rightarrow 196 \rightarrow 96 \rightarrow 32$) with layer norm [8] and LeakyRELU activation in between. The transposed convolution for temporal upsampling contains 4 1D transposed convolutional layers with stride=2 and kernel-size=3, padding=1.

4.2.3 Multi-Subject ECoG-to-Audio

The proposed SwinT allows the ECoG decoder to take input with any electrode layout and electrode order. Therefore, instead of training subject-specific ECoG decoder, our study makes it possible for the ECoG decoder to be shared among subjects. Figure 4.3 demonstrates the pipeline for multi-subject ECoG decoder training. Given multiple subjects, a shared SwinT-based ECoG decoder generates speech parameters based on each subject’s ECoG signal and electrode location (electrodes’ MNI coordinates and region index). Reference loss is calculated between the ECoG predicted speech parameters and the speech parameters generated by the subject-specific speech encoder. Each subject’s predicted speech parameters are then fed into the corresponding subject-specific speech synthesizer to generate speech spectrogram. During inference, the ECoG signal and electrodes’ location are fed into ECoG decoder to generate speech parameters. The subject’s speech synthesizer then generates speech spectrogram from the predicted speech parameters. Separate region embeddings are learned for left and right brain hemisphere when there are right hemisphere and left hemisphere subjects in the training set.

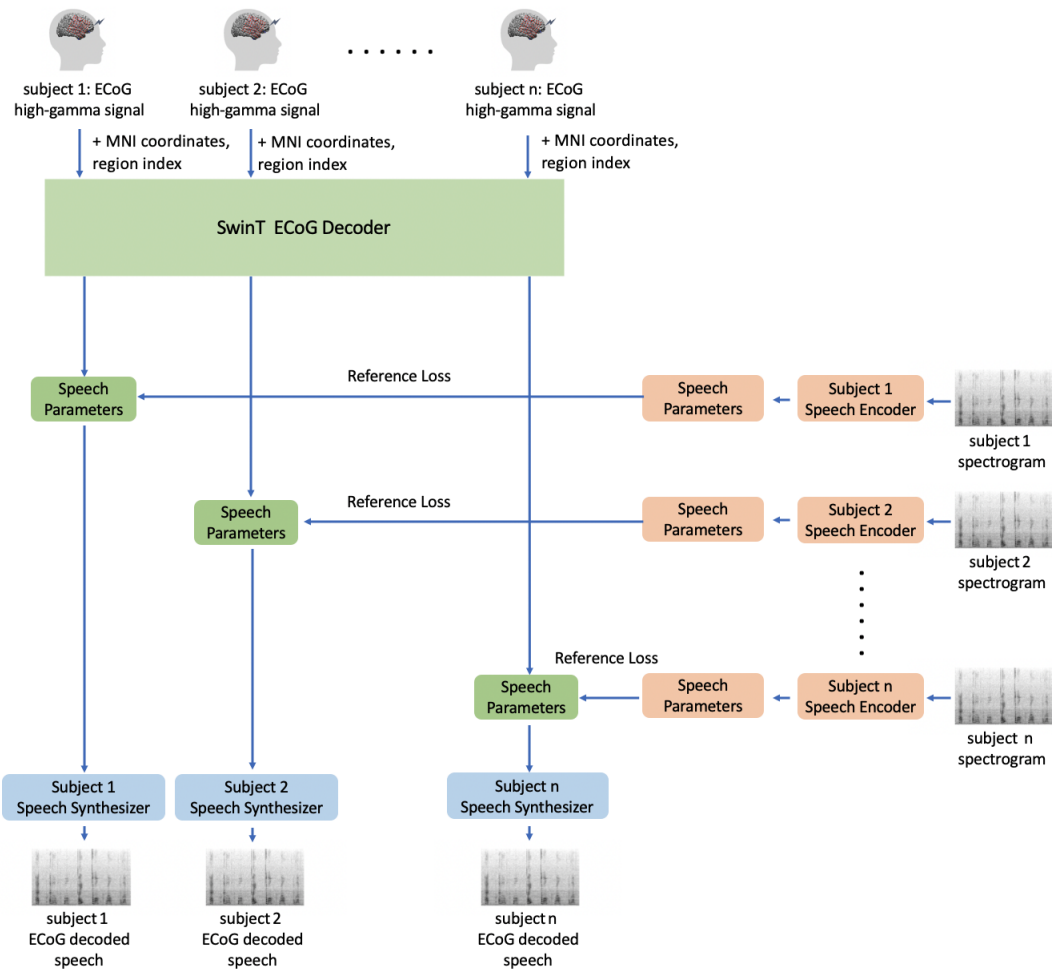


Figure 4.3: multiple-subject ECoG decoding training pipeline. Given multiple subjects, each subject’s ECoG signal and electrodes’ location information (MNI coordinates and ROI region index) are fed to a shared SwinT ECoG decoder to predict speech parameters. The predicted speech parameters are supervised by the speech parameters generated by subject-specific speech encoder from ground-truth speech spectrogram. The each subject’s predicted speech parameters are then fed into the corresponding subject-specific speech synthesizer to generate speech spectrogram.

4.2.4 Training of Speech Encoder and Speech Synthesizer

The audio-to-audio training of the speech encoder and learnable parameters of the speech synthesizer follows our previous work [28]. In summary, we train the speech encoder and speech synthesizer by letting them finish an audio-to-audio auto-encoding task, illustrated in (step 1) of Figure 4.1.

As detailed in [28], we supervise the training with multiple loss terms. The loss L_{a2a} consists of modified multi-scale spectral loss L_{MSS} , Short-Time Objective Intelligibility (STOI) loss L_{STOI} and supervision loss $L_{supervision}$, shown in the equation 4.4.

$$L_{a2a} = L_{MSS} + \lambda_1 L_{STOI} + \lambda_2 L_{supervision} \quad (4.4)$$

L_{MSS} is inspired by [41]. It supervises speech reconstruction by measuring the distance between the ground truth spectrogram and the reconstructed spectrogram in both linear scale and mel-frequency scale. L_{STOI} measures the intelligibility of reconstructed speech based on the STOI+ metric [49]. As higher STOI+ indicates better intelligibility, the L_{STOI} is defined as the negative of STOI+: $L_{STOI} = -STOI+$. Besides, additional supervision $L_{supervision}$ is applied to improve the accuracy of pitch f_0^t and formant frequencies $f_{i=1,2,3,4}^t$ prediction. The $L_{supervision}$ calculates the L2 distance between each predicted frequency and the corresponding frequency extracted by the Praat method [18]. The details of L_{MSS} , L_{STOI} and $L_{supervision}$ can be found in [28]. Following [28], λ_1 and λ_2 are set as 1.2 and 0.1, respectively. We use the speech synthesizer and speech encoder trained by our previous study [28]; training details can be found in [28].

4.2.5 Training of ECoG Decoder

Following [28], for the training of the ECoG decoder that predicts speech parameters from ECoG signals, we leverage two types of supervision to guide the training. Firstly, we train the ECoG decoder to generate speech parameters that match the parameters generated by the speech encoder. Besides, the ground truth speech spectrograms can act as additional supervision for the ECoG decoder, as the predicted speech parameters are converted to spectrograms by the speech synthesizer. For speech parameter based loss, we define reference loss $L_{reference}$ as equation 4.5:

$$L_{reference} = \sum_i \lambda_i \|\hat{C}_i^t - C_i^t\|_2^2 \quad (4.5)$$

$$i \in [f_0^t, f_1^t, \dots, f_6^t, a_1^t, \dots, a_6^t, f_u^t, b_u^t, a_u^t, \alpha^t, L^t] \quad (4.6)$$

$$L_{e2a} = L_{MSS} + \lambda_1 L_{STOI} + \lambda_2 L_{supervision} + \lambda_3 L_{reference} \quad (4.7)$$

where \hat{C}_i^t and C_i^t are speech parameters generated by the ECoG decoder and the speech encoder (as ground truth), respectively. We have 18 speech parameters defined in Section 4.2.1 and illustrated in the equation 4.6. We assign each speech parameter with individual weight λ_i , and the values are detailed in [28]. For spectrogram-based supervision, we follow the loss used in audio-to-audio training illustrated in equation 4.4. Therefore, the training loss for ECoG decoding is defined as equation 4.7 and λ_1 λ_2 and λ_3 are set as 1.2, 0.1 and 1.0.

Adam optimizer [71] with learning-rate= 5×10^{-4} , $\beta_1=0.9$ and $\beta_2=0.999$ is used to train the ECoG decoder. As mentioned in Section 4.3.1, following [28], randomly selected 50 out of 400 trials are used as the test set for each subject, and the rest of the data is used as the training set.

4.3 Results

4.3.1 ECOG Data Collection and Preprocessing

The study includes 43 native English-speaking subjects (20 males, 23 females) with refractory epilepsy (a disease involving seizures caused by abnormal electrical activity in brain cell communication). Details about speech and ECoG signals collection can be found in previous study [28]. In brief, at each trial, a subject was requested to speak a specific target word based on the stimuli provided by the care provider while their neural activity signals were recorded using ECoG electrodes. Each subject's trials were collected from 5 different tasks: (1) Auditory Repetition (repeating the word that the care provider has spoken), (2) Auditory Naming (naming the word based on definition that the care provider has spoken) (3) Sentence Completion (naming the last word to complete an sentence that the care provider has spoken) (4) Visual Reading (reading the written word shown by the care provider) (5) Picture Naming (naming the word based on a colored drawing shown by the care provider). Each task included 50 target words, each appearing once in the Auditory Naming and Sentence Completion and twice in each of the other tasks, leading to 400 trials of ECoG signal recording, and the average duration of word production among trials was 500ms.

In terms of the ECoG recording, detailed in [28], each of the 43 subjects has 8x8 electrodes with 10 mm spacing implanted to capture signals from the perisylvian cortex (male left hemisphere: 14 subjects; female left hemisphere: 13 subjects; male right hemisphere: 6 subjects; female right hemisphere: 10 subjects). The 8x8 electrodes are embedded in the perisylvian area as it is known to be the brain regions consisting of Broca's speech center and Wernicke's comprehension

center [34]. Besides 8x8 grid electrodes, each subject also has off-grid electrodes implanted, as strips outside of the 8x8 grid or as depth electrodes implanted under the surface of the brain, as shown in Figure 4.4. The experiments were approved by the Institutional Review Board of NYU Grossman School of Medicine, with written and oral consents collected from each participant. The ECoG arrays are FDA-approved. The high gamma component (70-150 Hz) was extracted, with electrodes with artifacts or interictal/epileptiform activity were excluded by setting their signal to 0. The details of the preprocessing can be found in [28]. This study also applies a Savitzky-Golay filter [106] with 3rd order polynomial and window size of 11 to further denoise the signal in the temporal dimension. Among the 400 trials of ECoG signals recorded from the five word production tasks, 350 trials were used for model training, and 50 trials were held out for testing (10 randomly selected trials were reserved for testing for each task).

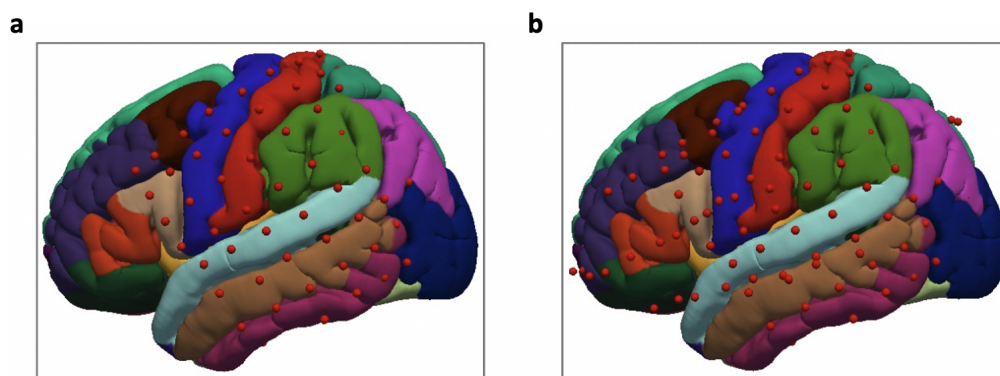


Figure 4.4: ECoG electrodes implanted on the human brain (a) 8x8 macro electrodes on the grid (b) both grid and off-grid electrodes.

4.3.2 Single-Subject Speech Decoding with Grid Electrodes

To compare our proposed grid-free SwinT with the ECoG decoders based on ResNet and 3D Swin transformer from the previous study [28], we firstly evaluated the SwinT trained with 64 ECoG electrodes on the grid for each subject individually. Following [28], we used two metrics to evaluate the speech decoding performance: 1) Pearson Correlation Coefficient (PCC) between the decoded spectrogram and the ground-truth spectrogram, and 2) STOI+ [49] that measures the intelligibility of the decoded speech (in the range between -1 to 1, higher STOI+ indicates better intelligibility). For each subject, we average PCC and STOI+ among all the test trials for model evaluation. As illustrated in Figure 4.5, the SwinT outperforms ResNet and 3D Swin transformer in terms of both CC and STOI+ for all subjects with only a few exceptions.. The superior performance of SwinT can also be demonstrated in Figure 4.6. Therefore, the results show that when tested on grid electrodes, although the SwinT does not have the inductive bias of grid layout and spatial locality in its architecture design, it can achieve better performance than grid-based ECoG decoders based on 3D Swin and ResNet. The results indicate that the SwinT is not only superior in its grid-free flexibility but also in the speech decoding performance.

4.3.3 Speech Decoding with Additional Off-Grid Electrodes

As the SwinT does not have spatial locality and assumption of grid input, unlike ECoG decoders based on ResNet or 3D Swin Transformer [28], the proposed SwinT can easily leverage off-grid electrodes to provide additional information for the speech decoding. In our study, for each subject, we selected off-grid electrodes

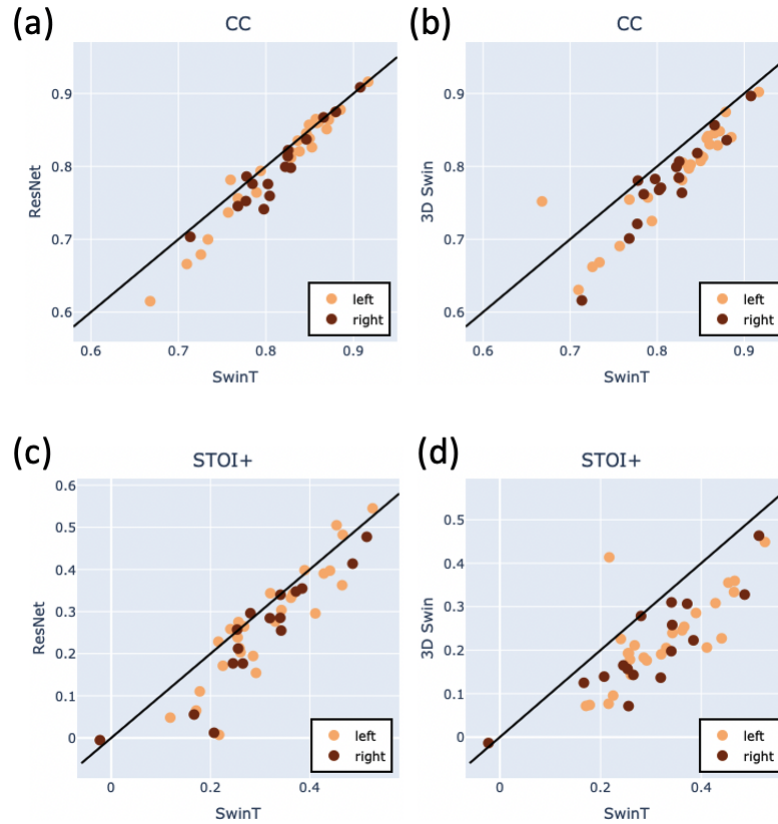


Figure 4.5: Comparison between baseline ECoG decoder and proposed SwinT when both trained and tested on grid electrodes: **(a)**: Comparison between ResNet and SwinT regarding PCC; **(b)**: Comparison between 3D Swin and SwinT regarding PCC; **(c)**: Comparison between ResNet and SwinT regarding STOI+; **(d)**: Comparison between 3D Swin and SwinT regarding STOI+. Each point indicates the speech decoding performance of a specific subject, with the x-axis as the performance of the SwinT and the y-axis as the performance of the baseline model. For (a)-(d), all sample points are below the diagonal line with only a few exceptions, indicating that the SwinT outperforms the two baseline models regarding both PCC and STOI+.

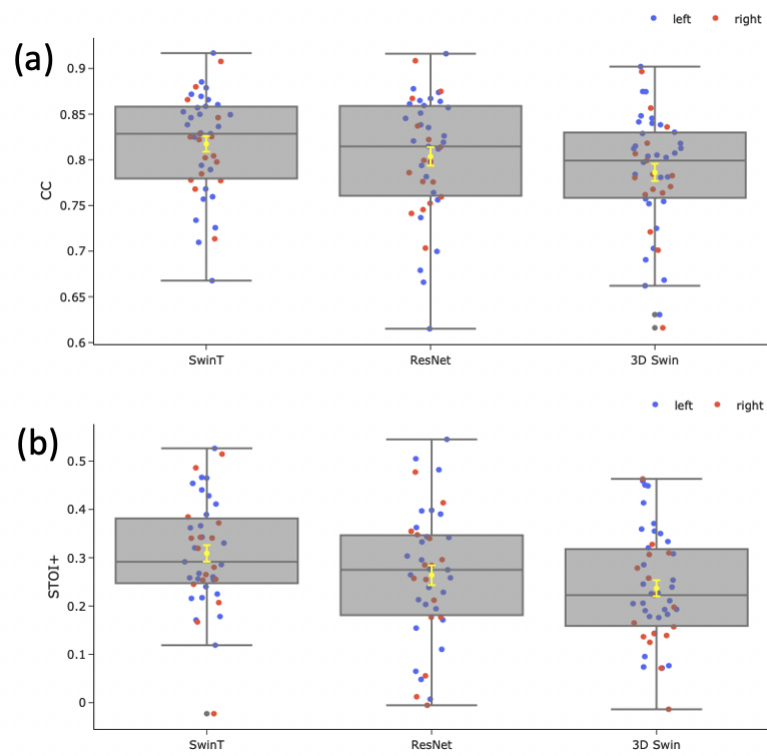


Figure 4.6: Comparison between baseline ECoG decoders (based on ResNet and 3D Swin Transformer) and the proposed SwinT ECoG decoder when models are trained and tested with grid electrodes for each subject individually: **(a)**: Comparison regarding PCC; **(b)**: Comparison regarding STOI+. The SwinT outperforms the baseline ECoG decoders based on ResNet and 3D Swin Transformer regarding both PCC and STOI+.

that have standard deviation of the signal greater than a subject-specific threshold. We applied the subject-specific threshold from a previous study [70]. We then trained the SwinT ECoG decoder with both 64 electrodes from the 8x8 grid and the selected off-grid electrodes for each subject, and calculate the PCC and STOI+ for the test trials. As there are 4 subjects that do not have any off-grid electrode fulfill the threshold requirement, we compared the models based on the remaining 39 subjects. Figure 4.7 illustrates the comparison between the SwinT ECoG decoders with and without the off-grid electrodes. The results demonstrate the superior performance of the SwinT ECoG decoder with off-grid electrodes as additional input. The grid-free flexibility of our proposed SwinT architecture can improve the performance of speech decoding by allowing the model to leverage the useful information in off-grid electrodes.

4.3.4 Speech Decoding Trained with Multiple-Subjects

As the proposed SwinT architecture does not require the electrodes to be arranged in a grid but relies on the electrode position in the brain anatomy, it is promising to handle the difference of electrode layout among different subjects and allows the ECoG decoder to be trained with multiple-subject data. To validate this idea, we trained a single SwinT ECoG decoder with 15 randomly selected male subjects with ECoG electrodes implanted in left or right brain hemisphere (left hemisphere: 4 subjects; right hemisphere: 11 subjects). As detailed in Section 4.2.3, subject-specific speech encoder and speech synthesizer are applied while the ECoG decoder is shared among subjects. We compared the SwinT trained with multiple subjects and the SwinT trained for single subject. We first evaluated the ECoG decoder on test trials of the 15 subjects. Illustrated in Figure 4.8, compared

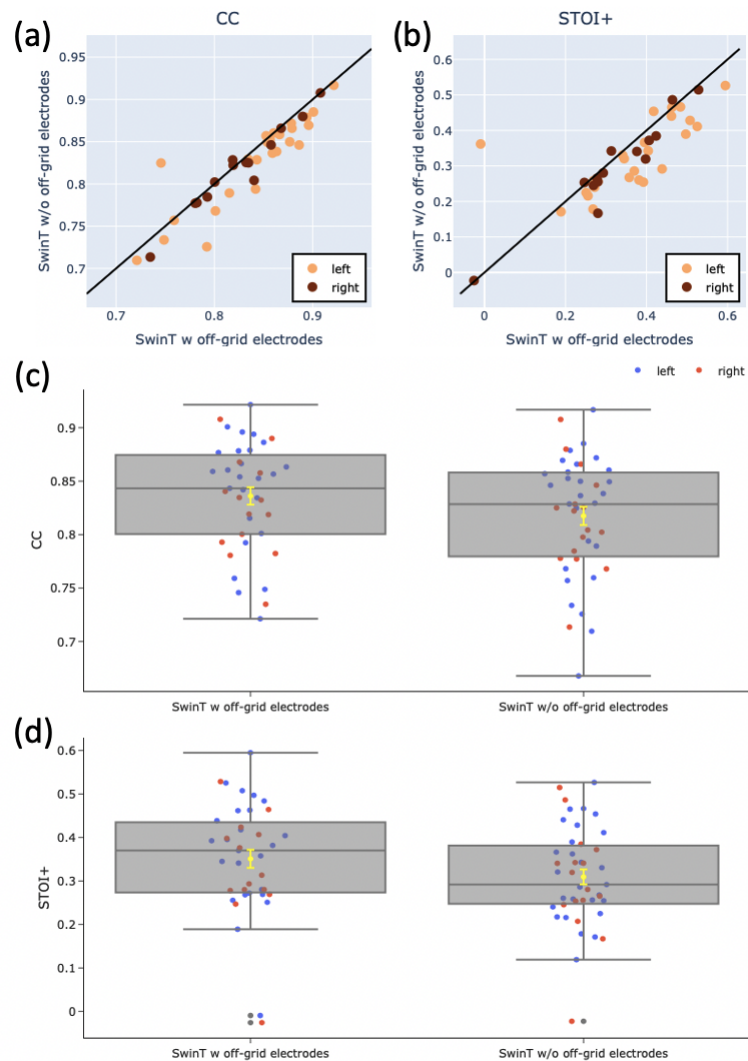


Figure 4.7: Comparison between SwinT ECoG decoders w and w/o off-grid electrodes, when trained on each subject individually. **(a) and (b)**: Comparison in PCC and STOI+, respectively. Each point indicates a subject, with x-axis being the performance of the SwinT with off-grid electrodes and y-axis being the performance of the Swin without off-grid electrodes; **(c) and (d)**: Comparison in PCC and STOI+ box plot. The results indicate the superior performance of the SwinT with off-grid electrodes.

with subject-specific SwinT, the SwinT shared among subjects achieved superior performance for most subjects and comparable performance for the rest.

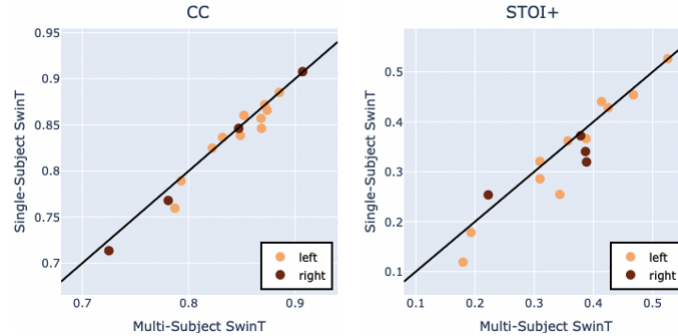


Figure 4.8: Comparison between the SwinT ECoG decoder trained with multiple (15) subject and the subject-specific SwinT. PCC and STOI+ were evaluated on test trials from the 15 subjects.

We also evaluated the multi-subject SwinT ECoG decoder on test trials of the subjects outside of the training set. We conducted 5-fold cross-validation separately for male and female subjects. Specifically, we partitioned subjects (with ECoG electrodes implanted in either left or right brain hemisphere) into five folds. Each time we use four folds of subjects to train a SwinT ECoG decoder and evaluate its ECoG decoding performance on the remaining one fold of subjects. The process is iterated to use every fold of subjects as the test subjects once. As shown in Figure 4.9, although the performance achieved on unseen subjects is significantly lower than the subject-specific models, the decoded speech still has a high pearson correlation with mean PCC=0.765. The results demonstrate the proposed SwinT ECoG decoder can achieve generalizability to subjects not in the training set. Additionally, as shown in Figure 4.10, compared with hemisphere-specific models, the SwinT ECoG decoder trained on both hemispheres can achieve comparable or slightly better performance when inferenced on unseen subjects.

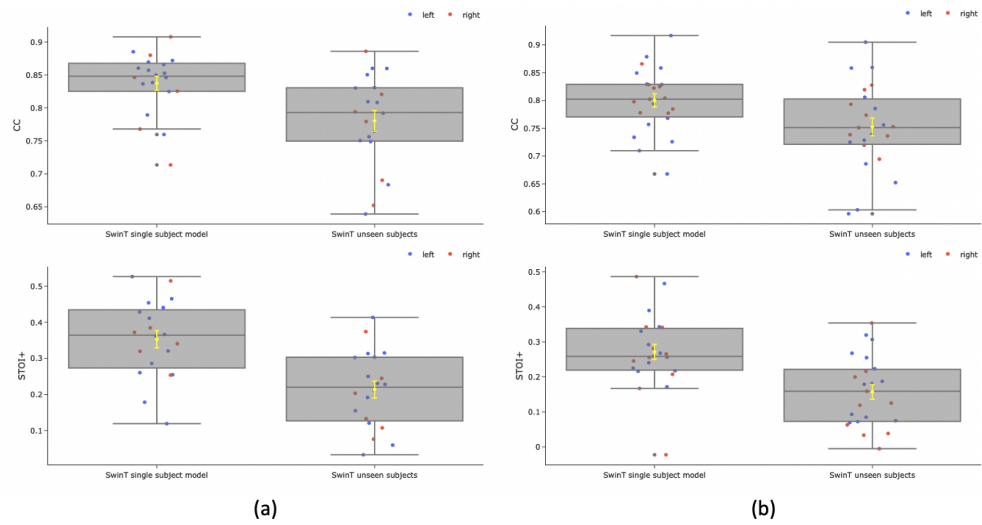


Figure 4.9: The performance of SwinT inferred on unseen subjects that are not in the training set. **(a)** cross-validation conducted on male subjects; **(b)** cross-validation conducted on female subjects. For each plot, the speech decoding performance when subjects are outside of the training subjects is shown on the right, and the performance of the SwinT decoder trained on each specific subject is shown on the left. The results demonstrate that the SwinT ECoG decoder can achieve generalizability to unseen subjects, as the performance achieved on unseen subjects is in a range with significant overlap with the performance of the single-subject models.

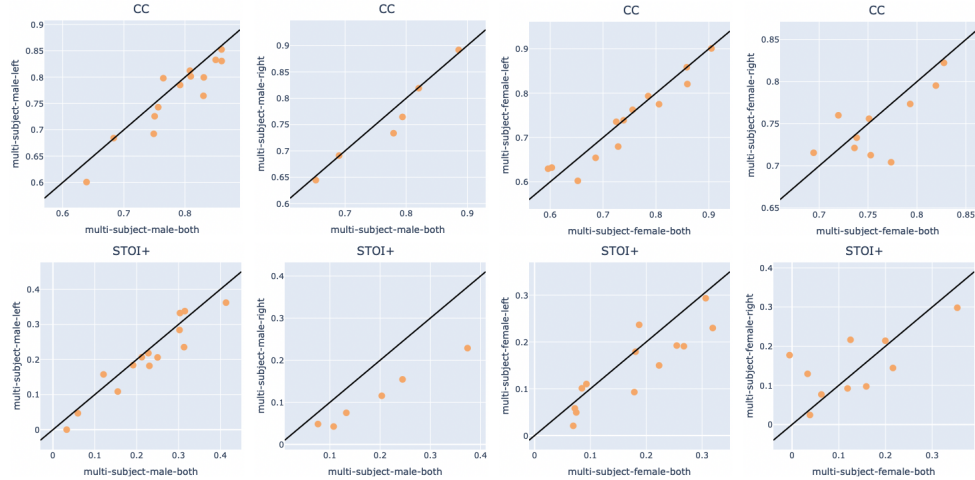


Figure 4.10: The comparison of speech decoding performance on unseen subjects between SwinT trained on one hemisphere and SwinT trained on both hemispheres. Models were trained separately for males and females. The results demonstrate that, compared with hemisphere-specific models, the SwinT ECoG decoder trained on both hemispheres can achieve comparable or slightly better performance when inferred on unseen subjects.

4.4 Discussion

This study proposes a new ECoG decoding architecture, SwinT, that does not have the grid-input assumption and can predict speech parameters from ECoG electrodes in any topological layout. The SwinT removes the grid-based operations in the Swin Transformers [77, 78, 79] (e.g., patch partition, patch merging, and local windowing) to make the model suitable for input in any layout. The grid-based operations are preserved for the temporal dimension as they provide the locality necessary to capture time-varying speech information. Besides, instead of relying on grid indexes to provide positional information about each electrode, the SwinT is fully anatomy-based as it relies only on electrodes’ position in the brain anatomy to generate relative positional bias for self-attention. Based on the 2-step training pipeline (Audio-to-Audio and ECoG-to-Audio training) and the speech

encoder/synthesizer proposed in our previous works [28, 122], the grid-free SwinT ECoG decoder can achieve better speech decoding performance compared to the previous grid-based ECoG decoders based on the ResNet and 3D Swin Transformer. Besides, as the SwinT does not require the electrodes to be arranged in a grid, it can further improve the speech decoding performance by leveraging off-grid electrodes as an additional source of information. Lastly, the proposed SwinT can also be trained with ECoG signals from multiple subjects and can achieve superior performance compared with the model trained for each subject individually. The multiple-subject SwinT can also achieve generalizability to new subjects out of the training set.

Our proposed SwinT ECoG decoder achieved superior performance than the ECoG decoders based on ResNet and 3D Swin Transformer from our previous studies [28, 122]. As illustrated in Figure 4.6, the SwinT achieved higher PCC and STOI+ (PCC: mean 0.817, median 0.828; STOI+: mean 0.309, median 0.292) than the ResNet (PCC: mean 0.804, median 0.815; STOI+: mean 0.264, median 0.275) and 3D Swin Transformer (PCC: mean: 0.785, median: 0.797; STOI+: mean 0.216, median 0.205) using the same 64 electrodes from the 8x8 ECoG grid. The results indicate that the SwinT can serve as a superior ECoG decoder compared with the ResNet and 3D Swin Transformer. Unlike ResNet and 3D Swin Transformer, the SwinT does not have spatial locality or grid input assumption. Therefore, the better performance achieved by the SwinT indicates that the SwinT can leverage the anatomical positional information provided by the MNI coordinates and region index to extract features that suit the speech decoding better.

Since the SwinT does not assume grid input format and can handle input electrodes with any topology, the SwinT can leverage off-grid electrodes that are

difficult to fit a grid (such as strip electrodes outside the grid or depth electrodes underneath the brain surface). Our results demonstrate that the proposed SwinT ECoG decoder architecture can leverage off-grid electrodes to improve speech decoding performance. As illustrated in Figure 4.7, for subjects with additional off-grid electrodes, the SwinT with additional off-grid electrodes achieved better PCC (mean 0.836, median 0.843) and STOI+ (mean 0.351, median 0.369) compared with the SwinT trained with grid electrodes only (PCC: mean 0.825, median 0.829; STOI+: mean 0.318, median 0.320). The superior results indicate that the neural activity recorded by the off-grid electrodes contains useful information correlated with the speech. The superior performance achieved by adding off-grid electrodes demonstrates the superiority of the SwinT’s grid-free and fully anatomy-based architecture.

The proposed SwinT achieved performance improvement when being trained with ECoG signals from multiple subjects (PCC: mean 0.837, median 0.849; STOI+: mean 0.352, median 0.378), compared with the SwinT trained for each subject individually (PCC: mean: 0.831, median: 0.846; STOI+: mean: 0.334, median: 0.340). The performance improvement can be explained by the more training samples and higher data diversity from the multiple subjects. Including multiple subjects for ECoG decoder training did not lead to performance improvement when we experimented with ResNet and 3D Swin Transformer. The success achieved by the SwinT can be attributed to its grid-free architecture. As the SwinT does not require grid layout of ECoG electrodes but the position in brain anatomy, the SwinT can better handle the differences in electrode placement between subjects. Besides, as illustrated in Figure 4.8, the SwinT can be successfully trained with subjects with ECoG implanted with left and right hemispheres. The success of the

left and right hemispheres co-training demonstrates the strong learning capacity of the SwinT. The two-hemisphere co-training also allows the ECoG decoder to fully leverage the whole dataset as we no longer need to train the model separately for each hemisphere.

Our study achieved generalizability to subjects outside of the training cohorts with the SwinT ECoG decoder trained with multiple subjects. Figure 4.9 shows that the speech decoding performance achieved on unseen subjects by multi-subject SwinT ECoG decoder is in a range with significant overlap with the results achieved by the subject-specific model. And the speech decoding performance is comparable between unseen subjects from the left and right hemispheres. The results indicate that the SwinT has successfully learned how to handle differences among subjects based on ECoG signals and the anatomical position of the electrodes. The result demonstrates the good prospect of ECoG-based speech decoding in real applications, as we can train a reliable decoder with other subjects and then directly deploy the model to the new subject.

There are several limitations of our study. Firstly, our speech decoding pipeline needs the audio synthesizer trained for each specific subject at the inference stage. Although the subject-specific audio synthesizer can help reconstruct the voice characteristics of the subject, it potentially makes the model not applicable to subjects with speech disabilities. To make our study work for paralyzed subjects, old speech recordings of the subjects can be used for the Audio-to-Audio training to get the speech synthesizer and speech encoder. Besides, the performance of subjects outside of the training cohorts is not consistently high. This could be potentially solved by including more subjects for the training when larger datasets become available. Lastly, our study focuses on word-level decoding. Sentence-level

study could be conducted in the future.

4.5 Conclusion and Contributions

In our study, we proposed a new neural network architecture named SwinT. As the ECoG decoder, the SwinT can predict speech parameters from ECoG signals and electrodes' position in the brain anatomy without requiring the electrodes to be arranged in a grid. The new ECoG decoder demonstrated superior speech decoding performance compared with baselines based on ResNet and 3D Swin Transformers. Besides, the grid-free architecture of the SwinT can also allow the model to leverage off-grid electrodes to help the speech decoding. The SwinT can be trained with multiple subjects with ECoG implanted in the left and right brain hemispheres. The multi-subject SwinT demonstrated better performance compared with the subject-specific models. Lastly, for the first time, our SwinT achieved generalizability to subjects outside of the training cohorts.

Chapter 5

Semi-Supervised Learning for ECoG Decoder based on Latent Decomposition

5.1 Introduction

Speech decoding based on electrocorticographic (ECoG) recordings can benefit patients with speech disabilities by helping them communicate with Brain-Computer Interface (BCI) [20, 83, 90, 108]. Recent studies have been exploring deep neural networks to build ECoG speech decoder and have achieved promising results [28, 122]. However, there are many challenges that remain to be solved to push the boundary of the ECoG decoder. Firstly, it is challenging to collect large-scale datasets in ECoG studies. The ECoG electrodes require invasive surgery to be

Junbo Chen is the co-main driver of this study. Acknowledgment to Xupeng Chen, Chenqian Le, Dr. Ran Wang, Dr. Amirhossein Khalilian-Gourtani, Prof. Adeen Flinker, and Prof. Yao Wang for their collaboration and advice.

implanted in the brain of patients. As the data collection requires invasive surgery, the datasets of ECoG recordings are limited. On the other hand, deep neural networks are data-demanding. For example, the Vision Transformer is data-hungry and difficult to train [121], making it challenging to apply the model to ECoG studies. Besides, the neural activity underlying speech production is very variable and complex. Even with a single subject speaking the same word, speech signals can have variation [13, 129]. Speech production also requires collaborations between regions corresponding to different functions, such as motor, auditory, and language processing [17, 96, 116]. The complex mechanism and high variation associated with the speech production make it challenging for the ECoG decoder to capture.

Recent studies in self-supervised learning can shed light on the ECoG research. Self-supervised learning has achieved significant performance improvement for deep neural networks [56, 57, 126], where the models can learn representation with good generalizability by learning to do pretasks without any ground-truth annotation. However, designing self-supervised learning for ECoG speech decoding is challenging. On the one hand, given the complexity of the speech production mechanism, it is unclear which pretask can lead to latent representation generalizable to speech decoding. On the other hand, the pretasks in many self-supervised learning studies heavily rely on data augmentation [27]. However, designing data augmentation for ECoG signals is not trivial and requires domain knowledge.

In our study, we propose a semi-supervised learning method to pretrain ECoG speech decoder. The pretraining method achieved superior speech decoding performance for the SwinT ECoG decoder (detailed in chapter 4). Our work is inspired by the SWAP-VAE [75] and its idea of decomposing the latent representation into content and style. The SWAP-VAE designs a novel pretraining framework that

uses reconstruction to decompose the neural activity of the primate target reaching task into "content" representation of semantic information (the direction of target) and "style" representation of dynamic information (the motion dynamics of the reaching process). In our study, we aim to simplify speech decoding by decomposing the complex neural activity of speech production into word-level semantics and trial-level dynamics, denoted as "word latent" and "trial latent". Compared with the study of SWAP-VAE [75], the ECoG signals in our study are more complex. Therefore, we propose to leverage word labels in our semi-supervised pretraining and combine the reconstruction pretasks in the original SWAP-VAE [75] with multiple other pretasks and regularization to guide the semantic and dynamics learning. Besides, the framework is carefully designed to prevent the collapsed solution that can fail the pretraining of the original SWAP-VAE [75].

5.2 Method

5.2.1 Pretraining Framework with Latent Decomposition

We propose a novel framework to pretrain the ECoG decoder with multiple pretasks and regularizations. In the original SWAP-VAE [75], two augmented views from the same sample of neural activity recording are considered as a positive pair that share the same content semantics. The reconstruction is applied as the pretask: given the two positive samples denoted as x_1 and x_2 with corresponding latent representation $[c_1, s_1]$ and $[c_2, s_2]$ (c denotes content and s denotes style), in addition to the reconstruction of x with corresponding latents, the contents are swapped between the two samples and the model is also trained to reconstruct x_1 with $[c_2, s_1]$ and x_2 with $[c_1, s_2]$. The L2 distance between the c_1 and c_2 is

applied to ensure the same content is shared by the two views. However, the content semantics defined as the common information among augmented views from the same sample is unclear in the setting of ECoG recordings. Besides, the reconstruction (with swap) can lead the model to collapse content c and compress all the information to the style s , making pretraining trivial. In our study, we leverage the word label and define content semantics as the word-level semantics (denoted as W) and style as the trial-level dynamics (denoted as T). To further help guide the model to decompose the word-level semantics and prevent the collapsed solution, we designed our pretraining framework as a combination of multiple pretasks and regularizations, illustrated in the Figure 5.1.

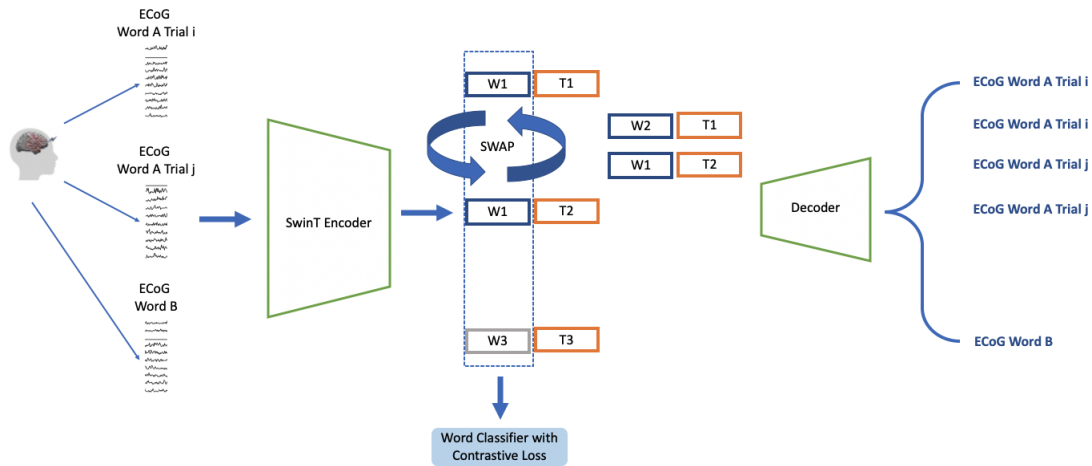


Figure 5.1: Framework for the ECoG decoder pretraining. At each iteration, three samples are input to the framework: two trials from the same word and one trial from a different word. The SwinT encoder extracts latent representation from each sample. The representation is divided into two parts with same dimension: word latent and trial latent, denoted as W and T . The latents of the original three latents go through decoder to reconstruct the corresponding signals. Besides, the content of the two same-word trials are swapped and used to predict the signals corresponding to the trial latent. The word latent are also fed to word classifier with contrastive loss.

Reconstruction with Swap: As shown in Figure 5.1, the pretraining requires

triplets of samples. At each iteration, two trials from the same word and one additional trial from a different word are input to the framework. The SwinT encoder extracts latent representation from each sample, with the architecture shown in Figure 5.2(a). The representation of each sample is divided into W and T with equal dimensions. Based on the decoder shown in the 5.2(b), each of the three latent representations is used to reconstruct the corresponding original ECoG signals. Inspired by [75], we also generate two additional latent by swapping the word latent between the two positive samples: $[W2, T1]$ and $[W1, T2]$. As they are different trials from the same word, they should have the same word semantics but different trial dynamics. Therefore, the $[W2, T1]$ and $[W1, T2]$ are used to generate the corresponding trials, which are ECoG signals of Word A Trial i and ECoG Word A Trial j in Figure 5.1.

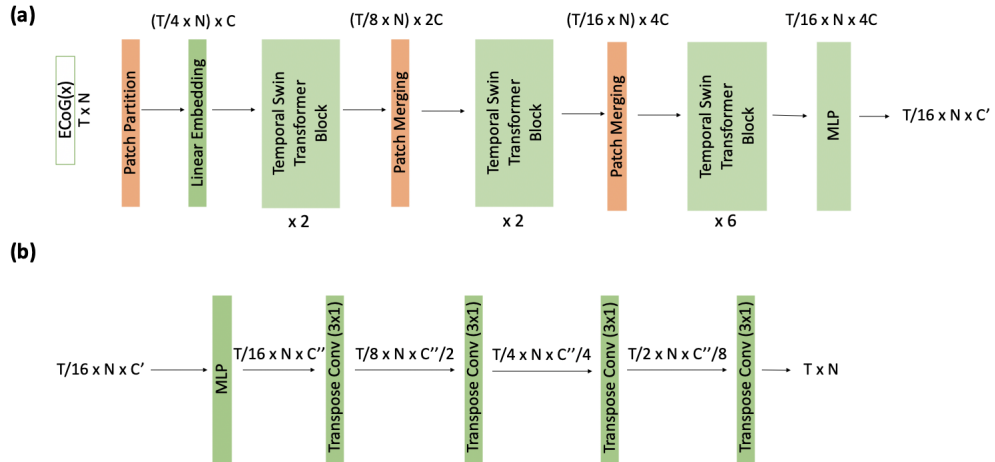


Figure 5.2: (a): SwinT Encoder, detailed in Section 4.2.2; (b): decoder consists of four transposed conv layers with kernel-size=3x1 to predict the signal reconstruction. $C = 96$, $C' = 64$, $C'' = 128$.

Word Prediction and Contrastive Learning: We further leverage the word label to guide the model to capture word semantics. As illustrated in Figure 5.1,

the word latents of the sample triplet are supervised by the contrastive loss defined in Section 5.2.3. In addition, two label-prediction tasks are also applied to guide the semantics learning: word prediction and positive-pair prediction. For word prediction, the prediction head illustrated in Figure 5.3 (a) is applied to predict the word label (50 words in total) from the W . For positive-pair prediction, we combine the W of every two samples and train a separate prediction head to predict if the two samples correspond to the same word, illustrated in Figure 5.3 (b).

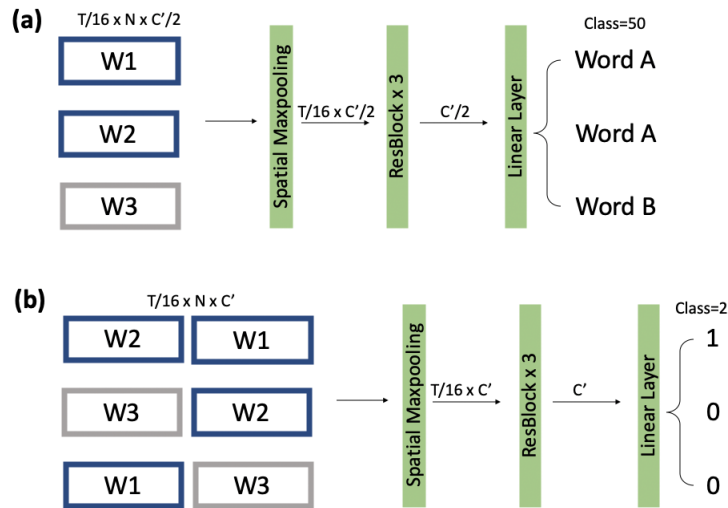


Figure 5.3: (a): Word classification prediction head; (b): If-same-word classification prediction head.

5.2.2 Data Augmentation

To further increase the generalizability of representations learned from the pretraining, we include a collection of data augmentations to the ECoG signals. The data augmentations included are: time jittering (shift signals in the time dimension by δ , with $8 < |\delta| < 16$), spatial dropout (randomly dropout electrode with $p = 0.2$), adding Salt and Pepper noise (magnitude=0.1 with $p=0.2$), adding

Gaussian noise (standard deviation is set as the standard deviation of the processed trial). We randomly select a subset of the four data augmentations at each time, with each method being included with $p = 0.5$.

5.2.3 Loss Functions and Training

In our pretraining framework, the framework requires three samples as an input triplet. The two positive samples corresponding to the same word are denoted as x_1 and x_2 . The additional negative sample from a different word is denoted as negative sample x_3 . The x denotes the ground-truth ECoG signal, and \hat{x} denotes the reconstructed signal. Besides, the \hat{x}_{swap_i} denotes the reconstructed signal from the swapped latent, with \hat{x}_{swap_i} referring to the trial latent is from the x_i . Given W_i and T_i indicating the word and trial latents of x_i , the \hat{x}_{swap_1} is reconstructed from $[W_2, T_1]$ and \hat{x}_{swap_2} is reconstructed from $[W_1, T_2]$.

Reconstruction Loss: The L2 distance between the ground-truth signals and the reconstructed signals is used as the reconstruction loss, denoted as the following:

$$L_{recon} = \sum_{i=1}^3 \|x_i - \hat{x}_i\|_2^2 + \sum_{i=1}^2 \|x_i - \hat{x}_{swap_i}\|_2^2 \quad (5.1)$$

However, the reconstruction loss can potentially lead to a collapsed solution. The model can simply compress all the information to the trial latent and collapse the word latent (e.g., word latent is always 0 regardless of the input signal) that contains no word information. To ensure the word latent contains useful information and prevent the collapsed solution, we add additional regularizations and supervision to the word latent.

Contrastive Loss/Triplet Loss: To prevent collapsed word latent and make it

represent word-level speech information, we design a simple triplet loss to supervise the word latent, defined as:

$$L_{triplet} = ||W_1 - W_2||_2^2 - ||W_1 - W_3||_2^2 - ||W_2 - W_3||_2^2 \quad (5.2)$$

We also tried other contrastive loss functions, such as InfoNCE [27] and hinged triplet loss [25], but the performance is not as good in our study. By optimizing the $L_{triplet}$, the word latents of trials from the same word will be pulled together, and trials from different words will be pushed away. The loss helps word latent to contain word-level semantic information, and pushing away latents from different words can also help prevent collapsed solutions.

Variance-Covariance Loss: To further prevent the word latent from collapsing and increase the information it contains, we apply the variance and covariance loss L_{VC} from [47] to the word latent, defined as:

$$L_{VC} = \lambda_C C(W) + \lambda_V V(W) \quad (5.3)$$

$$C(W) = \frac{1}{d} \sum_{i \neq j} Cov(W)_{i,j}^2 \quad (5.4)$$

$$V(W) = \frac{1}{d} \sum_{j=1}^d max(0, 1 - \sqrt{Var(W_{.,j})}) \quad (5.5)$$

The variance term $V(W)$ encourages each dimension to have high variance, ensuring that all dimensions are used in the latent. The covariance term $C(W)$ is to decorrelate the dimensions and encourage each dimension to capture different information. The λ_C and λ_V are set as 1 and 10 respectively.

Cross Entropy Loss: In our pretraining framework, we have two cross-entropy

loss terms: L_{word} for word classification and L_{same} for classifying if two latents are from trials corresponding to the same word.

Therefore, the loss we use for the pretraining is defined as:

$$L = L_{recon} + L_{triplet} + L_{VC} + L_{word} + L_{same} \quad (5.6)$$

The Adam optimizer [71] with learning-rate= 10^{-3} , $\beta_1=0.9$ and $\beta_2=0.999$ is used to train the ECoG decoder. After the pretraining, the weights of SwinT are used to initialize the SwinT ECoG decoder in the ECoG-to-audio training detailed in Chapter 4.

5.3 Results

We compared the speech decoding performance of the SwinT with and without the pretraining. For the pretrained SwinT, the SwinT is finetuned with loss function for ECoG decoding after the pretraining, detailed in Section 4.2.5. The experiments were conducted on the 43 subjects (each with 64 electrodes), detailed in Section 4.3.1. Speech decoding pipeline and SwinT ECoG decoder from Section 4.2.1 and 4.2.2 were used. For the pretrained SwinT, we evaluated the speech decoding performance of using both word latent W and trial latent T and the performances of using only one of them. The comparison of test STOI+ and PCC is shown in Figure 5.4, and the result shows the pretraining can lead to incremental performance gain. The pretrained SwinT using the three latents achieved mean PCC and STOI+ as $PCC_{W+T}=0.822$, $PCC_W=0.822$, $PCC_T=0.816$ and $STOI_{+W+T}=0.320$, $STOI_{+W}=0.317$, $STOI_{+T}=0.304$. Compared with the SwinT w/o pretraining (PCC: 0.817, STOI+: 0.309), both pretrained $W + T$ and

W only achieved performance improvement, and pretrained $W + T$ achieved the best result.

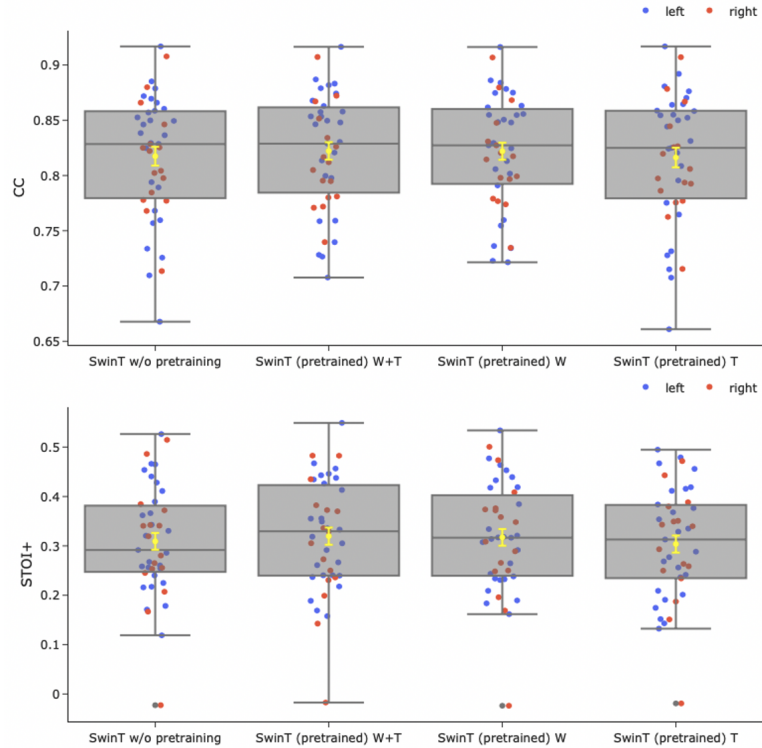


Figure 5.4: Performance comparison between ECoG decoder with and without pretraining. The performance of speech decoding are measured as PCC and STOI+. For pretrained SwinT ECoG decoder, we evaluated the speech decoding performances that rely on word+trial latent, word latent only, and trial latent only.

5.4 Discussion

This study propose a semi-supervised framework to pretrain ECoG speech decoder. By leveraging a collection of pretasks (reconstruction, swap reconstruction, word-level contrastive learning, word classification and same-word classification), the framework aims to train the ECoG decoder to decompose representation of

neural activity into word-level semantics and trial dynamics.

Multiple pretasks and regularizations are carefully selected and combined in the study to guide semantic learning and prevent collapsed solutions. Although our work is inspired by the SWAP-VAE [75], their framework does not suit the complex ECoG signals. Firstly, the model is pretrained to capture the shared information among positive samples as the semantics or content. The information shared by augmented views from the same ECoG signal does not correspond to any interpretable speech semantics. Therefore, we change the definition of positive samples as trials corresponding to the same word to capture the word-level semantics. Besides, the loss functions in the SWAP-VAE can lead to collapsed solutions as the model can simply set content latent to 0 and compress all information to the style, degenerating the pretraining to simple reconstruction. We observed this collapsed content (word semantics) in experiments. To regularize the word semantics from collapsing, we apply the variance and covariance loss from [47] and explicitly guide the word latent with contrastive loss and cross-entropy loss. The result shows that the collapsed solution is successfully prevented in our pretraining framework.

The ECoG decoder with the proposed semi-supervised pretraining achieved performance improvement compared with the ECoG decoder trained from scratch. The speech decoding based on both word semantics and trial dynamics achieved the best performance. Besides, speech decoding based on word semantics and trial dynamics individually can also achieve promising results. The results demonstrate the proposed pretraining is capable of learning representation generalizable to speech decoding, and validate the idea of simplifying the speech decoding task by decomposing the complex speech-related neural activity into word semantics and trial dynamics.

The current results show several limitations. Firstly, the performance increase from the pretraining is marginal. Secondly, the word semantics still needs improvement since the word semantics does not show good word clustering in test trials when visualized with T-SNE. Besides, the trial dynamics currently capture both speech dynamics and speech-nonrelated dynamics. Therefore, further disentanglement of the two dynamics is also needed. Lastly, the detailed interpretation of what types of semantics and dynamics is needed to shed light on the mechanism underlying speech production in the human brain.

5.5 Conclusion and Contributions

The study proposes a semi-supervised pretraining method to capture representation that can be leveraged to decode speech from ECoG recordings. The pretraining framework aims to simplify the complex neural activity correlated with speech production by decomposing the representation into word-level semantics and trial-level dynamics. The framework contains several pretasks, leveraging word labels as additional supervision to guide semantics learning during the pretraining. The results show that the pretraining can lead to performance improvement, which validates the idea that the disentanglement between semantics and dynamics can help speech decoding.

Chapter 6

Conclusion

This thesis leverages machine learning to study the human brain from two aspects: understanding the microstructure of the brain and designing a neural activity decoder to predict human speech.

For the investigation of brain microstructure, we focus on leveraging classification as means of identifying microstructural differences of the target cohorts. Specifically, we design models to classify the target cohorts based on multi-shell diffusion MRI and interpret the learned classifiers to pinpoint important diffusion metrics or brain regions for the classification tasks. In the study of microstructural differences of RHI, the classification pipeline with wrapper-based feature selection achieved promising results in classifying RHI subjects. The results support the notion that there are detectable white matter microstructure changes in the setting of RHI. The learned weights of the classifiers further reveal the influential diffusion metrics associated with RHI. The work serves as an example of methods that lead to a better understanding of the myriad of diffusion metrics as they relate to injury and disease. In the study of sex-related differences at the microscopic level, the

designed 3D CNN, 2D CNN, and Vision Transformer sex classifiers can achieve promising sex classification performance based on multiple volumetric diffusion metrics. The results demonstrate that the proposed MAE-based pretraining can lead to significant performance improvement to data-demanding ViT in the setting of multi-shell diffusion MRI. The occlusion analysis reveals white matter regions contribute most to sex-related differences and supports the idea of using distinctive neural networks to capture complementary information associated with sex-related differences. The results indicate that distinctive neural networks can capture complementary information regarding sex-related differences, and provide new insight supporting differences between male and female brain cellular-level tissue.

For decoding speech from human neural activity, we first propose a novel ECoG speech decoder, named SwinT. Instead of relying on any grid index, the SwinT leverages each electrode’s anatomical position and brain parcellation to decode human speech, enabling the model architecture to accommodate arbitrarily positioned electrodes. The proposed model achieved state-of-the-art performance based on the same grid electrodes used in the previous studies. It also achieved further performance increases by leveraging off-grid electrodes. More importantly, instead of relying on subject-specific ECoG decoders, our SwinT can be trained with ECoG signals from multiple subjects. The SwinT trained with multiple subjects not only achieved performance increase but also demonstrated generalizability to unseen subjects outside of the training set.

To further improve speech decoding, we propose a novel semi-supervised pre-training approach for the feature extraction part of the SwinT decoder. The study aims to simplify the complex neural activity associated with speech production by decomposing the latent representation into word-level semantics and

trial-level dynamics. The pretraining framework combines the pretasks of neural signal reconstruction and contrastive learning to guide the decomposition. Refining the pretrained network with the decoding loss led to improved speech decoding performance compared to training from scratch.

Bibliography

- [1] E. Adeli, Q. Zhao, N. M. Zahr, A. Goldstone, A. Pfefferbaum, E. V. Sullivan, and K. M. Pohl. Deep learning identifies morphological determinants of sex differences in the pre-adolescent brain. *NeuroImage*, 223:117293, 2020.
- [2] B. Ades-Aron, J. Veraart, P. Kochunov, S. McGuire, P. Sherman, E. Kellner, D. S. Novikov, and E. Fieremans. Evaluation of the accuracy and precision of the diffusion parameter estimation with gibbs and noise removal pipeline. *Neuroimage*, 183:532–543, 2018.
- [3] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.
- [4] G. K. Anumanchipalli, J. Chartier, and E. F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.
- [5] M. Asperholm, N. Högman, J. Rafi, and A. Herlitz. What did you do yesterday? a meta-analysis of sex differences in episodic memory. *Psychological bulletin*, 145(8):785, 2019.

- [6] M. Asperholm, L. Van Leuven, and A. Herlitz. Sex differences in episodic memory variance. *Frontiers in Psychology*, 11:613, 2020.
- [7] P. D. Asselin, Y. Gu, K. Merchant-Borna, B. Abar, D. W. Wright, X. Qiu, and J. J. Bazarian. Spatial regression analysis of mr diffusion reveals subject-specific white matter changes associated with repetitive head impacts in contact sports. *Scientific reports*, 10(1):13606, 2020.
- [8] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [9] J. F. Baizabal-Carvallo and J. Jankovic. Sex differences in patients with tourette syndrome. *CNS spectrums*, 28(2):205–211, 2023.
- [10] N. Barnea-Goraly, H. Kwon, V. Menon, S. Eliez, L. Lotspeich, and A. L. Reiss. White matter structure in autism: preliminary evidence from diffusion tensor imaging. *Biological psychiatry*, 55(3):323–326, 2004.
- [11] C. M. Baugh, J. M. Stamm, D. O. Riley, B. E. Gavett, M. E. Shenton, A. Lin, C. J. Nowinski, R. C. Cantu, A. C. McKee, and R. A. Stern. Chronic traumatic encephalopathy: neurodegeneration following repetitive concussive and subconcussive brain trauma. *Brain imaging and behavior*, 6:244–254, 2012.
- [12] J. J. Bazarian, T. Zhu, J. Zhong, D. Janigro, E. Rozen, A. Roberts, H. Javien, K. Merchant-Borna, B. Abar, and E. G. Blackman. Persistent, long-term cerebral white matter changes after sports-related repetitive head impacts. *PloS one*, 9(4):e94734, 2014.

- [13] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouviet, L. Fissore, P. Laface, A. Mertins, C. Ris, et al. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786, 2007.
- [14] M. Bergamino, E. G. Keeling, V. R. Mishra, A. M. Stokes, and R. R. Walsh. Assessing white matter pathology in early-stage parkinson disease using diffusion mri: A systematic review. *Frontiers in neurology*, 11:314, 2020.
- [15] Y. Bi, A. Abrol, Z. Fu, J. Chen, J. Liu, and V. Calhoun. Prediction of gender from longitudinal mri data via deep learning on adolescent data reveals unique patterns associated with brain structure and change over a two-year period. *Journal of Neuroscience Methods*, 384:109744, 2023.
- [16] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [17] S. C. Blank, S. K. Scott, K. Murphy, E. Warburton, and R. J. Wise. Speech production: Wernicke, broca and beyond. *Brain*, 125(8):1829–1838, 2002.
- [18] P. Boersma and V. Van Heuven. Speak and unspeak with praat. *Glott International*, 5(9/10):341–347, 2001.
- [19] S. P. Broglio, M. McCrea, T. McAllister, J. Harezlak, B. Katz, D. Hack, and B. Hainline. A national study on the effects of concussion in collegiate athletes and us military service academy members: the ncaa–dod concussion assessment, research and education (care) consortium structure and methods. *Sports medicine*, 47:1437–1451, 2017.

- [20] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther. Brain-computer interfaces for speech communication. *Speech communication*, 52(4):367–379, 2010.
- [21] B. Caplan, J. Bogner, L. Brenner, H. G. Belanger, R. D. Vanderploeg, and T. McAllister. Subconcussive blows to the head: a formative review of short-term clinical outcomes. *Journal of head trauma rehabilitation*, 31(3):159–166, 2016.
- [22] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [23] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski. Progress in speech decoding from the electrocorticogram. *Biomedical Engineering Letters*, 5:10–21, 2015.
- [24] R. A. Charlton, F. Schiavone, T. Barrick, R. Morris, and H. Markus. Diffusion tensor imaging detects age related white matter change over a 2 year follow-up which is associated with working memory decline. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(1):13–19, 2010.
- [25] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010.
- [26] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

- [27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [28] X. Chen, R. Wang, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker. A neural speech decoding framework leveraging deep learning and speech synthesis. *bioRxiv*, pages 2023–09, 2023.
- [29] B. C. Chiluba. Tackling disability of speech due to stroke: Perspectives from stroke caregivers of the university teaching hospital in zambia. *Indonesian Journal of Disability Studies*, 6(2):215–222, 2019.
- [30] S. Chung, J. Chen, T. Li, Y. Wang, and Y. Lui. Investigating brain white matter in football players with and without concussion using a biophysical model from multishell diffusion mri. *American Journal of Neuroradiology*, 43(6):823–828, 2022.
- [31] S. Chung, E. Fieremans, N. E. Kucukboyaci, X. Wang, C. J. Morton, D. S. Novikov, J. F. Rath, and Y. W. Lui. Working memory and brain tissue microstructure: white matter tract integrity based on multi-shell diffusion mri. *Scientific reports*, 8(1):3175, 2018.
- [32] J. A. Clayton and F. S. Collins. Policy: Nih to balance sex in cell and animal studies. *Nature*, 509(7500):282–283, 2014.
- [33] Q. Collier, J. Veraart, B. Jeurissen, A. J. den Dekker, and J. Sijbers. Iterative reweighted linear least squares for accurate, fast, and robust estimation of

- diffusion magnetic resonance parameters. *Magnetic resonance in medicine*, 73(6):2174–2184, 2015.
- [34] A. R. Damasio. Aphasia. *New England Journal of Medicine*, 326(8):531–539, 1992.
- [35] E. M. Davenport, K. Apkarian, C. T. Whitlow, J. E. Urban, J. H. Jensen, E. Szuch, M. A. Espeland, Y. Jung, D. A. Rosenbaum, G. A. Gioia, et al. Abnormalities in diffusional kurtosis metrics related to head impact exposure in a season of high school varsity football. *Journal of Neurotrauma*, 33(23):2133–2146, 2016.
- [36] S. De Santis, T. Granberg, R. Ouellette, C. A. Treaba, E. Herranz, Q. Fan, C. Mainero, and N. Toschi. Evidence of early microstructural white matter abnormalities in multiple sclerosis from multi-shell diffusion mri. *NeuroImage: Clinical*, 22:101699, 2019.
- [37] X. Ding, X. Zhang, J. Han, and G. Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11963–11975, 2022.
- [38] S. Dorfberger, E. Adi-Japha, and A. Karni. Sex differences in motor performance and motor learning in children and adolescents: an increasing male advantage in motor learning and consolidation phase gains. *Behavioural brain research*, 198(1):165–171, 2009.
- [39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image

- is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [40] S. J. Duff and E. Hampson. A sex difference on a novel spatial working memory task in humans. *Brain and cognition*, 47(3):470–493, 2001.
- [41] J. Engel, L. Hantrakul, C. Gu, and A. Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.
- [42] R. D. Fields. White matter in learning, cognition and psychiatric disorders. *Trends in neurosciences*, 31(7):361–370, 2008.
- [43] E. Fieremans, J. H. Jensen, and J. A. Helpert. White matter characterization with diffusional kurtosis imaging. *Neuroimage*, 58(1):177–188, 2011.
- [44] J. H. Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.
- [45] L. F. Gabler, S. H. Huddleston, N. Z. Dau, D. J. Lessley, K. B. Arbogast, X. Thompson, J. E. Resch, and J. R. Crandall. On-field performance of an instrumented mouthguard for detecting head impacts in american football. *Annals of biomedical engineering*, 48:2599–2612, 2020.
- [46] N. Gajawelli, Y. Lao, M. L. Apuzzo, R. Romano, C. Liu, S. Tsao, D. Hwang, B. Wilkins, N. Lepore, and M. Law. Neuroimaging changes in the brain in contact versus noncontact sport athletes using diffusion tensor imaging. *World neurosurgery*, 80(6):824–828, 2013.
- [47] Q. Garrido, L. Najman, and Y. Lecun. Self-supervised learning of split invariant equivariant representations. *arXiv preprint arXiv:2302.10283*, 2023.

- [48] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013.
- [49] S. Graetzer and C. Hopkins. Intelligibility prediction for speech mixed with white gaussian noise at low signal-to-noise ratios. *The Journal of the Acoustical Society of America*, 149(2):1346–1362, 2021.
- [50] D. Guneykaya, A. Ivanov, D. P. Hernandez, V. Haage, B. Wojtas, N. Meyer, M. Maricos, P. Jordan, A. Buonfiglioli, B. Gielniewski, et al. Transcriptional and translational differences of microglia from male and female brains. *Cell reports*, 24(10):2773–2783, 2018.
- [51] J. Han, Y. Fan, K. Zhou, K. Blomgren, and R. A. Harris. Uncovering sex differences of rodent microglia. *Journal of neuroinflammation*, 18(1):1–11, 2021.
- [52] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, and W. Zhang. Accurate screening of covid-19 using attention-based deep 3d multiple instance learning. *IEEE transactions on medical imaging*, 39(8):2584–2594, 2020.
- [53] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.
- [54] J. R. Harrison, S. Bhatia, Z. X. Tan, A. Mirza-Davies, H. Benkert, C. M.

- Tax, and D. K. Jones. Imaging alzheimer’s genetic risk using diffusion mri: A systematic review. *NeuroImage: Clinical*, 27:102359, 2020.
- [55] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [56] H. He, F. Zhang, S. Pieper, N. Makris, Y. Rathi, W. Wells, and L. J. O’Donnell. Model and predict age and sex in healthy subjects using brain white matter features: a deep learning approach. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.
- [57] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [58] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [59] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. *arXiv preprint arXiv:2003.11055*, 2020.
- [60] A. Herlitz, E. Airaksinen, and E. Nordström. Sex differences in episodic memory: the impact of verbal and visuospatial ability. *Neuropsychology*, 13(4):590, 1999.
- [61] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. Unet 3+: A full-scale connected unet for medical

- image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.
- [62] M. Ingalhalikar, A. Smith, D. Parker, T. D. Satterthwaite, M. A. Elliott, K. Ruparel, H. Hakonarson, R. E. Gur, R. C. Gur, and R. Verma. Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828, 2014.
- [63] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.
- [64] M. Jacobs and C. Ellis. Aphasianomics: estimating the economic burden of poststroke aphasia in the united states. *Aphasiology*, 37(1):25–38, 2023.
- [65] N. Jahanshad and P. M. Thompson. Multimodal neuroimaging of male and female brain structure in health and disease across the life span. *Journal of Neuroscience Research*, 95(1-2):371–379, 2017.
- [66] I. O. Jelescu, M. Zurek, K. V. Winters, J. Veraart, A. Rajaratnam, N. S. Kim, J. S. Babb, T. M. Shepherd, D. S. Novikov, S. G. Kim, et al. In vivo quantification of demyelination and recovery using compartment-specific diffusion mri metrics validated by electron microscopy. *Neuroimage*, 132:104–114, 2016.
- [67] R. A. Kanaan, M. Allin, M. Picchioni, G. J. Barker, E. Daly, S. S. Shergill, J. Woolley, and P. K. McGuire. Gender differences in white matter microstructure. *PloS one*, 7(6):e38272, 2012.

- [68] S. B. Kaufman. Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, 35(3):211–223, 2007.
- [69] E. Kellner, B. Dhital, V. G. Kiselev, and M. Reiser. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magnetic resonance in medicine*, 76(5):1574–1581, 2016.
- [70] A. Khalilian-Gourtani, R. Wang, X. Chen, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker. A corollary discharge circuit in human speech. *BioRxiv*, pages 2022–09, 2022.
- [71] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [72] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.
- [73] G. Krishna, Y. Han, C. Tran, M. Carnahan, and A. H. Tewfik. State-of-the-art speech recognition using eeg and towards decoding of speech spectrum from eeg. *arXiv preprint arXiv:1908.05743*, 2019.
- [74] H. Liu and H. Motoda. *Feature extraction, construction and selection: A data mining perspective*, volume 453. Springer Science & Business Media, 1998.
- [75] R. Liu, M. Azabou, M. Dabagia, C.-H. Lin, M. Gheshlaghi Azar, K. Hengen, M. Valko, and E. Dyer. Drop, swap, and generate: A self-supervised approach for generating neural activity. *Advances in neural information processing systems*, 34:10587–10599, 2021.

- [76] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34:23818–23830, 2021.
- [77] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.
- [78] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [79] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.
- [80] P. Lolekha, K. Phanthumchinda, and R. Bhidayasiri. Prevalence and risk factors of parkinson’s disease in retired thai traditional boxers. *Movement Disorders*, 25(12):1895–1901, 2010.
- [81] M. Lotze, M. Domin, F. H. Gerlach, C. Gaser, E. Lueders, C. O. Schmidt, and N. Neumann. Novel findings from 2,838 adult brains on sex differences in gray matter brain volume. *Scientific reports*, 9(1):1671, 2019.
- [82] E. Luders, A. W. Toga, and P. M. Thompson. Why size matters: differences in brain volume account for apparent sex differences in callosal anatomy: the sexual dimorphism of the corpus callosum. *Neuroimage*, 84:820–824, 2014.

- [83] S. Luo, Q. Rabbani, and N. E. Crone. Brain-computer interface: applications to speech decoding and synthesis to augment communication. *Neurotherapeutics*, 19(1):263–273, 2022.
- [84] J. G. Makin, D. A. Moses, and E. F. Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582, 2020.
- [85] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, J. Lancaster, et al. A probabilistic atlas of the human brain: theory and rationale for its development. *Neuroimage*, 2(2):89–101, 1995.
- [86] T. W. McAllister, J. C. Ford, L. A. Flashman, A. Maerlender, R. M. Greenwald, J. G. Beckwith, R. P. Bolander, T. D. Tosteson, J. H. Turco, R. Raman, et al. Effect of head impacts on diffusivity measures in a cohort of collegiate contact sport athletes. *Neurology*, 82(1):63–69, 2014.
- [87] P. H. Montenegro, M. L. Alosco, B. M. Martin, D. H. Daneshvar, J. Mez, C. E. Chaisson, C. J. Nowinski, R. Au, A. C. McKee, R. C. Cantu, et al. Cumulative head impact exposure predicts later-life depression, apathy, executive dysfunction, and cognitive impairment in former high school and college football players. *Journal of neurotrauma*, 34(2):328–340, 2017.
- [88] P. Moreno-Briseño, R. Díaz, A. Campos-Romo, and J. Fernandez-Ruiz. Sex-related differences in motor learning and performance. *Behavioral and brain functions*, 6(1):1–4, 2010.
- [89] S. Mori, K. Oishi, H. Jiang, L. Jiang, X. Li, K. Akhter, K. Hua, A. V. Faria, A. Mahmood, R. Woods, et al. Stereotaxic white matter atlas based on

- diffusion tensor imaging in an icbm template. *Neuroimage*, 40(2):570–582, 2008.
- [90] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications*, 10(1):3096, 2019.
- [91] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [92] L. E. Nicholas and R. H. Brookshire. Comprehension of spoken narrative discourse by adults with aphasia, right-hemisphere brain damage, or traumatic brain injury. *American Journal of Speech-Language Pathology*, 4(3):69–81, 1995.
- [93] Y. Nonomura, M. Yasumoto, R. Yoshimura, K. Haraguchi, S. Ito, T. Akashi, and I. Ohashi. Relationship between bone marrow cellularity and apparent diffusion coefficient. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 13(5):757–760, 2001.
- [94] D. S. Novikov, E. Fieremans, S. N. Jespersen, and V. G. Kiselev. Quantifying brain microstructure with diffusion mri: Theory and parameter estimation. *NMR in Biomedicine*, 32(4):e3998, 2019.
- [95] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn:

- Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [96] W. Penfield and L. Roberts. *Speech and brain mechanisms*, volume 62. Princeton University Press, 2014.
- [97] L. Picco, M. Subramaniam, E. Abdin, J. A. Vaingankar, and S. A. Chong. Gender differences in major depressive disorder: findings from the singapore mental health study. *Singapore medical journal*, 58(11):649, 2017.
- [98] T. A. Pigott. Anxiety disorders in women. *Psychiatric Clinics*, 26(3):621–672, 2003.
- [99] E. M. Prager. Addressing sex as a biological variable, 2017.
- [100] N. F. Ramsey, E. Salari, E. J. Aarnoutse, M. J. Vansteensel, M. G. Bleichner, and Z. Freudenburg. Decoding spoken phonemes from sensorimotor cortex with high-density ecog grids. *Neuroimage*, 180:301–311, 2018.
- [101] A. Rezaei and L. C. Wu. Automated soccer head impact exposure tracking using video and deep learning. *Scientific reports*, 12(1):9282, 2022.
- [102] S. J. Ritchie, S. R. Cox, X. Shen, M. V. Lombardo, L. M. Reus, C. Alloza, M. A. Harris, H. L. Alderson, S. Hunter, E. Neilson, et al. Sex differences in the adult human brain: evidence from 5216 uk biobank participants. *Cerebral cortex*, 28(8):2959–2975, 2018.
- [103] A. N. Ruigrok, G. Salimi-Khorshidi, M.-C. Lai, S. Baron-Cohen, M. V. Lombardo, R. J. Tait, and J. Suckling. A meta-analysis of sex differences

- in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39:34–50, 2014.
- [104] S. G. Ryman, M. P. van den Heuvel, R. A. Yeo, A. Caprihan, J. Carrasco, A. A. Vakhtin, R. A. Flores, C. Wertz, and R. E. Jung. Sex differences in the relationship between white matter connectivity and creativity. *NeuroImage*, 101:380–389, 2014.
- [105] T. D. Satterthwaite, D. H. Wolf, D. R. Roalf, K. Ruparel, G. Erus, S. Vandekar, E. D. Gennatas, M. A. Elliott, A. Smith, H. Hakonarson, et al. Linked sex differences in cognition and functional connectivity in youth. *Cerebral cortex*, 25(9):2383–2394, 2015.
- [106] R. W. Schafer. What is a savitzky-golay filter?[lecture notes]. *IEEE Signal processing magazine*, 28(4):111–117, 2011.
- [107] D. K. Schneider, R. Galloway, J. J. Bazarian, J. A. Diekfuss, J. Dudley, J. L. Leach, R. Mannix, T. M. Talavage, W. Yuan, and G. D. Myer. Diffusion tensor imaging in athletes sustaining repetitive head impacts: a systematic review of prospective studies. *Journal of neurotrauma*, 36(20):2831–2849, 2019.
- [108] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg. Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2257–2271, 2017.
- [109] K. K. Seunarine, J. D. Clayden, S. Jentschke, M. Munoz, J. M. Cooper, M. J. Chadwick, T. Banks, F. Vargha-Khadem, and C. A. Clark. Sexual

- dimorphism in white matter developmental trajectories using tract-based spatial statistics. *Brain connectivity*, 6(1):37–47, 2016.
- [110] S. M. Smith, M. Jenkinson, H. Johansen-Berg, D. Rueckert, T. E. Nichols, C. E. Mackay, K. E. Watkins, O. Ciccarelli, M. Z. Cader, P. M. Matthews, et al. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31(4):1487–1505, 2006.
- [111] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004.
- [112] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- [113] F. Szczepankiewicz, S. Lasič, D. van Westen, P. C. Sundgren, E. Englund, C.-F. Westin, F. Ståhlberg, J. Lätt, D. Topgaard, and M. Nilsson. Quantification of microscopic diffusion anisotropy disentangles effects of orientation dispersion from microstructure: applications in healthy volunteers and in brain tumors. *Neuroimage*, 104:241–252, 2015.
- [114] M. TalavageThomas, A. NaumanEric, L. BreedloveEvan, E. DyeAnne, E. MorigakiKatherine, J. LeverenzLarry, et al. Functionally-detected cognitive impairment in high school football players without clinically-diagnosed concussion. *Journal of neurotrauma*, 2014.
- [115] R. Thomas, A. M. O’Connor, and S. Ashley. Speech and language disorders

- in patients with high grade glioma and its influence on prognosis. *Journal of neuro-oncology*, 23:265–270, 1995.
- [116] V. L. Towle, H.-A. Yoon, M. Castelle, J. C. Edgar, N. M. Biassou, D. M. Frim, J.-P. Spire, and M. H. Kohrman. Ecog gamma activity during a language task: differentiating expressive and receptive speech areas. *Brain*, 131(8):2013–2027, 2008.
- [117] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.
- [118] J. Veraart, E. Fieremans, and D. S. Novikov. Diffusion mri noise mapping using random matrix theory. *Magnetic resonance in medicine*, 76(5):1582–1593, 2016.
- [119] S. B. Vos, D. K. Jones, B. Jeurissen, M. A. Viergever, and A. Leemans. The influence of complex white matter architecture on the mean diffusivity in diffusion tensor mri of the human brain. *Neuroimage*, 59(3):2208–2216, 2012.
- [120] D. Voyer, S. D. Voyer, and J. Saint-Aubin. Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic bulletin & review*, 24:307–334, 2017.
- [121] P. Wang, X. Wang, H. Luo, J. Zhou, Z. Zhou, F. Wang, H. Li, and R. Jin. Scaled relu matters for training vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2495–2503, 2022.

- [122] R. Wang, X. Chen, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker. Distributed feedforward and feedback cortical processing supports human speech production. *Proceedings of the National Academy of Sciences*, 120(42):e2300255120, 2023.
- [123] D. M. Werling and D. H. Geschwind. Sex differences in autism spectrum disorders. *Current opinion in neurology*, 26(2):146, 2013.
- [124] X. Wu, I. I. Kirov, O. Gonen, Y. Ge, R. I. Grossman, and Y. W. Lui. Mr imaging applications in mild traumatic brain injury: an imaging update. *Radiology*, 279(3):693–707, 2016.
- [125] Y. Wu and K. He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [126] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.
- [127] T. Xu, X. Cai, Y. Wang, X. Wang, S. Chung, E. Fieremans, J. Rath, S. Flanagan, and Y. W. Lui. Identification of relevant diffusion mri metrics impacting cognitive functions using a novel feature selection method. In *2019 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–6. IEEE, 2019.
- [128] H. W. Yeung, S. Luz, S. R. Cox, C. R. Buchanan, H. C. Whalley, and K. M. Smith. Pipeline comparisons of convolutional neural networks for structural connectomes: predicting sex across 3,152 participants. In *2020 42nd Annual*

International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 1692–1695. IEEE, 2020.

- [129] P. Zelinka, M. Sigmund, and J. Schimmel. Impact of vocal effort variability on automatic speech recognition. *Speech Communication*, 54(6):732–742, 2012.
- [130] Y. Zhang, N. Schuff, A.-T. Du, H. J. Rosen, J. H. Kramer, M. L. Gorno-Tempini, B. L. Miller, and M. W. Weiner. White matter damage in frontotemporal dementia and alzheimer’s disease measured by diffusion mri. *Brain*, 132(9):2579–2592, 2009.
- [131] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan. A deep learning model integrating fcnn and crfs for brain tumor segmentation. *Medical image analysis*, 43:98–111, 2018.

Publication List

Junbo Chen*, Vara Lakshmi Bayanagari*, Sohae Chung, Yao Wang, Yvonne W. Lui, “*Deep Learning with Diffusion MRI as in vivo Microscope Reveals Sex-related Differences in Human White Matter Microstructure*” (*equal contribution) (submitted to Scientific Report)

Junbo Chen, Sohae Chung, Tianhao Li, Els Fieremans, Dmitry S Novikov, Yao Wang, Yvonne W Lui, “*Identifying relevant diffusion MRI microstructure biomarkers relating to exposure to repeated head impacts in contact sport athletes.*” in The Neuroradiology Journal, doi: <https://doi.org/10.1177/19714009231177396>.

Sohaе Chung, **Junbo Chen**, Tianhao Li, Yao Wang and Yvonne W Lui, “*Investigating Brain White Matter in Football Players with and without Concussion Using a Biophysical Model from Multi-shell Diffusion MRI.*” in American Journal of Neuroradiology, doi: <https://doi.org/10.3174/ajnr.A7522>.

Sohaе Chung, **Junbo Chen**, Tianhao Li, Yao Wang, Yvonne Lui. “*White Matter Microstructural Alterations In Contact-Sport Athletes With And Without Concussion: A Multi-Shell Diffusion Study*”, International Society for Magnetic Resonance in Medicine (ISMRM), 2021.

Junbo Chen, Sohae Chung, Joseph Rath, Els Fieremans, Yao Wang, Yvonne Lui. “*Predicting Visual-Motor Functioning in Patients with Mild Traumatic Brain Injury from Multi-Shell Diffusion MRI using Gradient Boosting Tree*”, Annual Meeting of Radiological Society of North America (RSNA), 2020.

Junbo Chen*, Jeffrey Mao*, Cassandra Thiel, Yao Wang. “*iWaste: Video-*

Based Medical Waste Detection and Classification". International Engineering in Medicine and Biology Conference (EMBC), IEEE, 2020. (*equal contribution; oral session)