Design and Development of Dyslexia Data Consortium

Roshan Bhandari
Clemson University, rbhanda@clemson.edu

Follow this and additional works at: https://tigerprints.clemson.edu/all_theses

Recommended Citation
https://tigerprints.clemson.edu/all_theses/3582

This Thesis is brought to you for free and open access by the Theses at TigerPrints. It has been accepted for inclusion in All Theses by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.
Abstract

Dyslexia is a neurodevelopmental reading disability and is supposed to be affecting about 5-15 percent of the population in the United States alone. However, neuroimaging studies in dyslexia research involve relatively small sample sizes, thus limiting inference and the application of novel methods. Besides, the lack of standards among datasets shared makes the datasets useless and also raises questions about the privacy and security of individual subjects involved in the research. Hence, it is essential to develop a data-sharing platform that solves all of these issues. In this thesis, we develop and describe of the platform Dyslexia Data Consortium. The overarching goal of this project is to advance our understanding of a disorder that has significant academic, social, and economic impacts on children, their families, and society. In this platform, researchers can upload and share dyslexia datasets for collaboration. Furthermore, a deep learning-enabled data quality check ensures that the data shared has all the features needed for study and ensures that the subjects' privacy is protected. Thus, researchers can access shared data to address fundamental questions about dyslexia, replicate findings, apply new methods, and educate the next generation of dyslexia researchers. Moreover, the platform generates secondary datasets from the shared datasets and provides them to researchers. These secondary datasets can be helpful to determine how much previous findings replicate in their samples. In addition, brain volume estimation and correlation analysis help researchers get answers to questions related to dyslexia.
Acknowledgments

I am grateful to Dr. James Wang for his inspiration for this work and his patience, advice, and suggestions throughout my master’s degree. Dr. Mark Eckert, professor at College of Medicine, Medical University of South Carolina, and Dr. Kenneth Vaden, Research Assistant Professor at the Medical University of South Carolina, also deserve a round of applause for their collaboration and guidance throughout the project development. I also want to thank Dr. Mark Smotherman, Dr. Amy Apon, Dr. Nina Hubig, and the graduate student coordinator Adam Rollins for their guidance and support during my master’s studies at the Clemson University. I would also like to thank Dr. Feng Luo and Dr. Pradip Srimani, my thesis committee members, for their guidance and support. I would also like to thank Akash Indani from Clemson University for his help in CNN model training process. Chuck Cook, the systems programmer at the School of Computing, also deserves a big thanks for the technical support in setting up the computing environment for the dyslexia data consortium. Finally, I would like to thank Xizhou Feng from the CCIT for his help in setting up the data processing pipeline in Palmetto.
Table of Contents

Title Page ................................................................. i
Abstract ................................................................. ii
Acknowledgments ......................................................... iii
List of Tables ............................................................ vi
List of Figures ........................................................... vii
1 Introduction ............................................................ 1
  1.1 Thesis Organization ............................................... 2
2 Related Works .......................................................... 3
3 System Design and Architecture ...................................... 5
  3.1 System Architecture ............................................... 5
  3.2 Database Diagram ................................................. 5
  3.3 Data Processing And API ......................................... 6
  3.4 Tools and Technologies .......................................... 8
4 Data Sharing and Storage ............................................. 10
  4.1 User Interface For Registration .................................. 11
  4.2 User Interface for Dashboard .................................... 12
  4.3 Data Upload ....................................................... 12
  4.4 User Interface For Upload ....................................... 13
  4.5 Data Download ................................................... 15
  4.6 User Interface For Data Download ............................... 15
  4.7 BIDS Compliance .................................................. 16
5 SECONDARY DATASET GENERATION .................................. 19
  5.1 FMRI Analysis ..................................................... 19
  5.2 Secondary Image generation steps in block diagram .......... 20
  5.3 Different Types Of Secondary Images ............................ 21
  5.4 Native Space Gray/White Matter ................................ 21
  5.5 Forward/Inverse Warped Images ................................ 22
  5.6 Modulated and Normalized Grey/White Matter ................. 22
  5.7 Modified Native Grey/White Matter .............................. 22
  5.8 Normalized Bias Field Corrected Image ........................ 23
  5.9 Surface Data .................................................... 23
6 Parallel Data Processing In Palmetto ............................... 25
List of Tables

6.1 API format for jobs, upload, download and status api . . . . . . . . . . . . . . . . . 27
6.2 API format for correlation and secondary dataset parameters update api . . . . . . 28
6.3 Comparision of number of datasets that can be processed in Palmetto V Single Server 30
# List of Figures

3.1 System Architecture Diagram .................................................. 6  
3.2 Database Design ...................................................................... 7  
3.3 Data Flow in Dyslexia Data Consortium ..................................... 8  
3.4 Jobs API sample ...................................................................... 8  
4.1 Signup Page ............................................................................ 11  
4.2 Dashboard Page ....................................................................... 12  
4.3 Upload Page ............................................................................ 14  
4.4 Compulsory (first) stage of mapping ......................................... 15  
4.5 Download Page ....................................................................... 16  
4.6 BIDS representation of a sample dataset taken from bids.neuroimaging.io ................................................. 17  
5.1 Data Processing Steps ................................................................. 20  
6.1 Parallel processing using Palmetto step wise ............................... 29  
6.2 Performance Analysis Palmetto V Consortium ............................ 30  
7.1 Data Quality Steps .................................................................. 33  
7.2 Data Quality Check UI ............................................................... 34  
7.3 CNN Model Training Process .................................................... 35  
7.4 Prediction Workflow ................................................................. 36  
7.5 Sample 2D representation of 3D image uploaded ....................... 38  
7.6 Co-variance calculation for image similarity ............................... 38  
8.1 Brain Volume Estimation Steps .................................................. 41  
8.2 User Interface for Brain Metrics ................................................ 42  
9.1 DeId toolbox share page updated .............................................. 45  
9.2 DeID Integration Workflow ....................................................... 46
Chapter 1

Introduction

Dyslexia is a neurodevelopmental reading disability that affects people of different age groups and is more common in children. Research studies have estimated that about 5-15% of the population have dyslexia\cite{14}\cite{32}\cite{37}. Many reasons such as family history and genetics, premature birth, exposure to different kinds of drugs during pregnancy, etc., cause dyslexia. Dyslexia relates to different brain functions and brain regions. Different types of research are going on in the dyslexia domain. These researches explore various topics such as the relation between different brain sections and the cause of dyslexia, etc.

The human brain is complex and consists of different regions. Researchers have been trying to understand various aspects of brain regions. The frontal brain is the most significant part of the brain and is concerned with controlling speech and reasoning. The parietal lobe controls understanding of what we hear and read. Modern research revolves around understanding brain functions and the impact on these brain regions. To understand it fully and develop proper remedies and treatments, we need to leverage the advancements in data science. A platform where researchers can share data quickly and efficiently will enhance collaboration and continuation of a research beyond the original goal, ultimately bringing innovation.

This thesis introduces a web platform where researchers can upload and share dyslexia datasets to create a collaborative environment that empowers modern research. Researchers can access data to address fundamental questions about dyslexia, replicate findings, apply new methods, and educate the next generation of dyslexia researchers. It produces secondary data sets such as gray matter images, modulated gray matter images, modulated normalized gray matter images, etc.,
from the uploaded 3D MRI images and helps researchers in the holistic understanding of different subjects in dyslexia research. Researchers can use the secondary datasets generated by the platform to replicate the results in their samples. The platform also provides brain (gray matter) volumes, supports data quality analysis to ensure shared images do not contain facial features, and ensures no brain tissue voxels are removed.

1.1 Thesis Organization

This thesis consists of ten chapters. First, it introduces dyslexia, and related research works in the domain. Then it revolves around the system design and analysis. After system design and analysis, it focuses on different system functions and their implementation from chapter four till chapter eight. These chapters cover data sharing and collaboration, data storage, bids format for standardization, secondary datasets generation, parallel processing using palmetto, data quality using deep learning models, and brain volume estimation using the cat12 toolbox. The integration of popular data de-identification tool, DeId toolbox[39] in the Dyslexia Data Consortium is discussed in chapter 9. The final chapters discuss, analyze the results and conclude the thesis.
Chapter 2

Related Works

Dyslexia affects many persons, especially children. Studying neuroimaging datasets can help us understand dyslexia better. They help us understand the brain and how the brain functions in healthy and unhealthy conditions. Different researches are conducted to better understand brain using neuroimaging studies[36][21]. These researches generate a huge volume of datasets that researchers can share to advance the studies further beyond the original scope. However, the lack of a proper data sharing and management platform makes the data generated disseminate only to a small group of researchers. Dyslexia study typically involves relatively small sample sizes that limit inference and the application of novel approaches that can require large sample sizes. Data sharing and collaboration is the only way to solve the problem.

The advancement of computational and storage technologies has made it possible for digital innovations to happen frequently. These technologies are transforming the world and have made it possible for researchers to collaborate and develop novel methods. Data sharing platforms make it possible to share datasets and promote open access to data. They make it possible to share knowledge by decreasing the data management efforts. Some platforms go beyond and make it possible to share computational resources too. Data sharing promotes transparency, open access, and collaboration, ultimately increasing the efficiency of all stakeholders. As such, funding agencies such as NSF are concerned about having unrestricted access to data and results of research[6].

There are quite a few platforms that help users share neuroimaging datasets. One of such is neurovault.org[18]. Researchers can use neurovault to store and share data produced from MRI and PET studies[18]. Another is brainmap.org[24]. BrainMap stores published datasets of func-
tional and structural neuroimaging experiments. These datasets can exist in MNI space and be coordinate-based. These platforms have data of different fMRI studies and are extensively used by researchers for different research. Another platform, called ANIMA, provides a platform for sharing results of fMRI studies.

All these platforms serve their specific purpose; however, none of these are specially designed for dyslexia. They do not have automated methods to compute morphology measures predictive of reading disability. They also do not have an automated process to audit the data. Researchers have to audit their datasets manually before uploading them to those platforms. There is also heterogeneity in data collected from different sources. The heterogeneity in data sources makes it very hard for researchers to collect accurate morphology measures for images collected from various sources. Moreover, standard specifications like BIDS lack appropriate tools to help standardize the datasets collected from multiple sources. A tool that provides a user interface that lets user map different data sources as anatomical, diffusion, functional, etc., is needed for the BIDS generator to solve the problem. Also, open access to data means concerns such as privacy would arise. Additional rules and compliance standards like Health Insurance Privacy and Portability Act (HIPPA) must be followed while sharing data. Tools like the DeId toolbox removes identifications associated with datasets and makes it possible to share demographic and behavioral datasets. Integrating the DeID toolbox into the data-sharing platform can help user deidentify their datasets before sharing them publicly.

To address the issues mentioned above, we design and develop Dyslexia Data Consortium to help researchers share datasets and collaborate on research. This platform establishes new methods to compute and provide researchers brain morphology measures. The computation of these morphology measures is possible by exploiting massive parallel computation with Clemson University’s Palmetto supercomputer. The shared datasets and computed morphological metrics such as datasets (gray matter image, white matter image, brain volumes, etc.) can be made publicly available for download. The platform also develops and integrates deep learning models to perform data quality. The models detect the facial features (identification information) in data and the absence of brain voxels tissues, thus helping researchers automate the manual work in data quality. Integration of the DeID toolbox with the platform makes it easy for researchers to deidentify their datasets if identity information is present. Moreover, datasets uploaded to the platform are stored and shared in BIDS compliant format, resolving the issues related to the data heterogeneity.
Chapter 3

System Design and Architecture

3.1 System Architecture

Dyslexia Data Consortium is a platform intended to help collaborators share data sets related to dyslexia research. The platform is built using Python and Django web framework and follows client/server and Model View Template (MVT) architecture. The Models represent the database and object-oriented representation of the database tables. The Templates represent the HTML design and the user interfaces. Finally, the Views is the central logic to route data to and from the template and the database. Figure 3.1 represents the architecture diagram of the system. On the client-side, a user sends a request to the Dyslexia Data Consortium using the browser of their choice. Then, the request is forwarded to the Django URL handler by the server’s apache web server. The URL forwards the request to the corresponding View, where application logic resides. The View then manipulates the models for data and also renders the template to the user.

3.2 Database Diagram

Figure 3.2 shows the database design of the Dyslexia Data Consortium. This database includes six tables currently. More tables might be needed if new functions or data are introduced in the future. Users table saves basic information such as the username and active status of a user. UserProfile table stores other details of the user signup form such as name, organization, the purpose. Table MRIDataset holds information such as the name of the dataset, brain file (3D
Figure 3.1: System Architecture Diagram

Mri image, behavior file, and a foreign key to table "Users" to track users uploading the datasets. MRIDataset also stores information about the dataset and prediction scores obtained from deep learning models for brain cut prediction and face removed predictions. Table MRIDatasetCorrelation has information such as correlation with the template image, max correlation among dataset pool, etc. It also has a foreign key to MRIDataset to track the dataset for which correlation is calculated. SecondaryDatasetParameters holds data such as brain volume, cortical thickness, etc., obtained after secondary dataset generation. It also has a foreign key to the table MRIDataset to keep track of the datasets for which secondary datasets were generated. Finally, the table DatasetDownloadRequest contains information related to dataset download requests made by a user and has foreign keys to table MRIDataset and Users table.

3.3 Data Processing And API

Figure 3.3 shows the data flow diagram of the Dyslexia Data Consortium system. There are three layers in which data processing happens—first, the user system layer, where all the requests made by users exist. The second layer is the API layer. In this layer, for all the datasets uploaded, dataset processing jobs exist. A data processing job contains information such as username, dataset link for download, etc., in JSON format and can be used by a script in Palmetto supercomputer for parallel processing. The third layer is the parallel processing layer. In this layer, multiple tasks
are processed using parallel processing. Works such as secondary dataset generation, correlation calculation, brain volume estimation, etc., are done in this layer. After the processing, the status gets updated, and the outcomes and results are sent to the Dyslexia Data Consortium database using the API. Figure 3.4 shows the sample data and request format for the Jobs API.
3.4 Tools and Technologies

3.4.1 Web Application

Dyslexia Data Consortium web application runs on a dell server powered with Ubuntu Linux OS and connected to palmetto supercomputer using API for parallel processing. Users can use any browser to access the system at dyslexia.computing.clemson.edu publicly. Apache webserver is used
to host server-side web applications. HTML5, CSS3, Javascript are used as frontend technology to build the user interface. To save the records of users, and datasets, and other processed data information such as prediction, brain volumes, etc., PostgreSQL is used. Python version 3.6 and Django Framework are used to develop the web application.

3.4.2 API Development

The data processing pipelines come with the capability of parallel processing on multiple machines. For parallel processing, the system exposes an API. These processing jobs use these API endpoints to pull datasets processing details. These jobs also use these APIs to track the job status and upload back the processing outcomes. Django REST Framework (DRF) is used to build the API. Authentication using OAuth makes the API secured.

3.4.3 Data Processing

The data processing pipeline uses Matlab, SPM12, and cat12 toolbox. SPM12 provides different functionalities to analyze images of functional neuroimaging experiments. CAT12 toolbox is an extension of SPM12 and provides features such as voxel-based morphometry, surface-based morphometry, and deformation-based morphometry. We use SPM12 and CAT12 toolbox functionalities for secondary image generation, correlation analysis, and brain volume estimation. Python scripts are used to call the Matlab scripts, which call the SPM12 and CAT12 utilities.

3.4.4 Deep Learning and Prediction

Dyslexia Data Consortium uses two deep learning models to assist users in data quality checks. The first model predicts if the dataset uploaded by a user has facial features present or not. In contrast, the second model predicts if brain tissue voxels are removed or not during the de-identification process. The models are written in python and the TensorFlow library.
Chapter 4

Data Sharing and Storage

Researchers around the world have conducted various studies in the neuroimaging domain. These researches lead to the accumulation of a large amount of dataset. These datasets can be beneficial to future researchers, and they can inspire innovation and novel methods that can prove beneficial to the improvement of human health. Jean et al. highlight several reasons for sharing neuroimaging dataset. Some of them are to expedite our understanding of the brain, develop better research, and reduce research costs. Many researchers are sharing their datasets actively. Some of them are available in neurovault.org, some in brainmap.org and others are in platforms like Kaggle but there are a lot of research datasets that are not shared. Also, the datasets that are shared do not have uniformity in format. Besides, these platforms do not provide any computational advantage in generating other brain morphological datasets or estimating brain volume. They also do not automate human efforts involved in the data quality, nor do they try to address privacy protection and security. With these issues, the proper dataset for research is always lacking, which has been a critical factor in slowing down the efforts of future researchers.

Dyslexia Data Consortium aims to bridge the gap by creating a data-sharing platform where researchers can collaborate for dyslexia research. It provides an accessible web platform for researchers to join and upload datasets in a standard way defined by BIDS specifications. In addition, it uses massively parallel computational technology to produce other brain morphological datasets and analyses. These morphological measures can be used to replicate research findings as

1 https://www.kaggle.com/search?q=fmri
2 https://github.com/zh1peng/open-access-fMRI-database
well as further the research goals.

4.1 User Interface For Registration

![Signup Page](image)

Figure 4.1: Signup Page

All users must be registered in the Dyslexia Data Consortium to share and download datasets. Figure 4.1 is a registration page. Users can go to this page through a "signup" link available on the home page. The user can enter the name, username, email address, organization, and purpose to register on this page. A user account in the Dyslexia Data Consortium needs to be approved by the admin. Once a user fills in details and clicks the signup button, the admin is notified about the registration. Admin approves his account, then a notification email about the approval is sent to the user, and he can start using the system.
4.2 User Interface for Dashboard

Figure 4.2 is an image of the dashboard in the Dyslexia Data Consortium. It highlights different features available on the website. Using the upload icon, users can go to the upload page. Using the download button, users can go to the download page to download the dataset. The data quality page takes users to a data quality page where users can view the predictions related to face present or not and brain cut. On another page, brain metrics take users to a page where users can view OFC volume, STS volume, etc. Create Post page is a website management tool from where admin user can create post, announcement, etc. In the "view post" page, a user can see a listing of different pages created by the admin. Admin menu takes a user to a page, where the user can do user management, profile management, email management, etc. The profile page takes a user to a page where a user can edit his personal information. Finally, with the change password menu, the user can change his/her password.

Figure 4.2: Dashboard Page

4.3 Data Upload

One of the primary purposes of developing a Dyslexia Data Consortium is to promote data sharing and research collaboration. To share datasets, users can select the upload menu from the dashboard after login. The upload process and the user interface are amicable, and users can upload
images without much training about the Dyslexia Data Consortium. Users can either upload a single image in nii format or a zip or tar file of many MRI images. They can also upload a zip file containing multiple zip files. Users will need to specify the image type (functional, anatomical, or diffusion) to make the data sharing bids compliant. Once the image type is selected, appropriate follow-up questions appear. Zip or tar file after the upload completes gets extracted and stored in BIDS format. Users can also upload a file generated directly from the DeID toolbox. Users also need to specify if the dataset is shared publicly or available to a specific group of users or is private during the upload process. After the upload is complete, the page redirects to the data mapping page. Users need to map the behavior datasets, if any, to standard fields in the Dyslexia Data Consortium.

4.4 User Interface For Upload

Figure 4.3 shows the upload page in the Dyslexia Data Consortium. On this page, the user needs to set the name for the dataset they want to upload. Users can also attach the new data to an existing dataset by selecting the dataset name from the drop-down list. Next, users need to enter a short description of the dataset in the description field. A yes/no button is there to choose if the uploaded dataset is a DeId toolbox output or not. Users also need to select the image type of the dataset uploaded. The image type field is used to store the dataset in BIDS-compliant format. Option to choose a dataset, metadata related to the dataset, behavior dataset, and metadata related to the dataset is available. Finally, users need to select if they have permission to upload the dataset on the consortium and with whom they want to share the dataset. During the upload process, users can select who can access the datasets. Making a dataset "private" means it is visible only to the owner of the datasets, whereas making it "public" means all researchers on the consortium can access it. Users can also add limited access to the dataset they upload. Limited access datasets are visible and available for download only to a small group of users.

After the user clicks the upload button on the upload page, they are redirected to a page to map their dataset to the standard format of Dyslexia Data Consortium. This mapping process has one mandatory and four optional stages. Users need to select basic mapping like id, age, sex variable in the first compulsory stage. Then, in the optional four steps, they will need to map different experiments and their data to the standard fields available in the Dyslexia Data Consortium, if
available in their dataset. This step is necessary to bring uniformity in the datasets uploaded by different users. Figure 4.4 shows the mapping page user interface.
4.5 Data Download

Researchers collaborate on Dyslexia Data Consortium by sharing datasets. The link to the dataset download page is available in the dashboard. Once a user clicks and reaches the download page, he can see a list of datasets. He can select them and request them for download by clicking the download button. On the sidebar, users can click the filters to choose the dataset they want to download. Once a user selects datasets and requests the download, a download request is submitted. The job processor processes these download requests on a first-come, first-serve basis. All the requested datasets are zipped and mailed to the user when it is ready.

4.6 User Interface For Data Download

Figure 4.5 is a download page. Users can go to the download page by clicking the download option in the dashboard. The download page lists the dataset available to the user. All public and datasets shared with the users are available for download on the page. To download the dataset, users need to select datasets. Users can also apply different kinds of filters to download the dataset. Once the selection is complete, the user needs to click the download button. Once the download button is clicked, a dataset download request is sent to the server. The server will process the request and zip datasets available for download on a first-request, first-serve basis and send a link to the user.
user via email. Users can use this link and download the requested dataset.

![Download Page](image.png)

Figure 4.5: Download Page

### 4.7 BIDS Compliance

Magnetic resonance imaging (MRI) is beneficial for the study of the brain. Many types of research involving the brain adopt this technique. MRI focuses on image data, but the structure of the image files was not well defined. As a result, different researchers store their MRI images and associated behavior data or secondary datasets in different formats. That’s where BIDS comes into the picture. It stands for Brain Imaging Data Structure. It consists of standard practices for storage, organization, and description of MRI datasets [17].

BIDS describes a simple and easy way to organize neuroimaging and behavioral data. It
is easy to follow and can standardize many neuroimaging experiments. BIDS defines some aspects as required; meanwhile, some elements are optional but regulated. For example, it states that a dataset must have at least one subject and gives a format to save a T1-weighted image [17]. BIDS also provides naming conventions for structural, diffusion, functional brain images, and associated behavioral data. According to it, to represent different subjects, we must create separate folders. The folder names must have a unique identifier, for example, sub-01, sub-02, etc. Under the subject folder, there should be multiple folders to place anatomical, functional, or diffusion images. It also states that a dataset owner must create a tab-separated file called participants.tsv to store details of participants. Figure 4.6 from bids.neuroimaging.io shows a BIDS representation of a sample dataset on the left side of the image on the right side.

Dyslexia Data Consortium uses BIDS conventions to organize the data collected from different researchers. UI and the backend code is developed to ensure BIDS format for data storage. The upload page asks a series of questions related to image types to the user. First, users need to select the image type, i.e., if the image is functional, anatomical, or diffusion. If it is anatomical, they also need to choose if it is a T1W or T2W. If the image type is functional, the user should also enter a task description (rest or other). If it is diffusion, then a task label needs to be entered. Once a user clicks the upload button, the data goes through a BIDS parser that will save the file in the BIDS format, as shown in figure 4.6.
One of the key benefits of following the BIDS convention is that it standardizes data sharing. Furthermore, it can also act as a convenient method to de-identify the MRI images in the dataset uploaded. The identification information in the dataset name, which the researchers often use in their local system, gets removed by bids compliant subject names(sub-1, sub-2, etc.).
Chapter 5

SECONDARY DATASET GENERATION

This chapter describes different techniques used to generate secondary datasets and gives an overview of secondary datasets generated by the Dyslexia Data Consortium. These secondary datasets may be helpful for different kinds of research related to the human brain and dyslexia.

5.1 FMRI Analysis

Magnetic Resonance Imaging (MRI) is used in many research to collect images of internal structures of the human body. MRI uses different pulse sequences to visualize different types of contrast visualized using magnetic fields, radio waves, and a computer to initiate the sequence and construct the images. Functional Magnetic Resonance Imaging (fMRI) can be used to measure blood oxygen level changes in the human brain. Researchers can also use fMRI to measure changes in blood flow in the brain when a person is doing different activities [33]. It measures the changes in blood oxygenation by detecting magnetic signals resulting from the hemoglobin under conditions such as sleeping, resting, running, etc. Hemoglobin has diamagnetic behavior when it is oxygenated but paramagnetic behavior when deoxygenated. fMRI makes it possible to measure and discriminate the cerebral spinal fluid, white matter, and gray matter. During an fMRI session, a patient performs different kinds of tasks that increase or decrease the supply of oxygenated blood to a particular part
of the brain.

To process the 3D MRI images, we use SPM12 (Statistical Parametric Mapping) and CAT12 (Computational Anatomy Toolbox) software. It is open-source software available for free. Researchers can use SPM12 to analyze brain imaging sequences such as time-series data of one subject, etc. CAT12 toolbox allows us to do voxel-based morphometry analysis and surface-based analysis. Using python scripts to call the Matlab scripts consisting of spm12 and cat12 functionalities, we generate analysis reports of the 3D fMRI images.

5.2 Secondary Image generation steps in block diagram

![Data Processing Steps](image)

Figure 5.1: Data Processing Steps

Figure 5.1 represents series of data processing steps in the generation of a secondary dataset. Once a user uploads a 3D MRI image on the Dyslexia Data Consortium, the system creates a secondary dataset generation job. Then it makes it available to the Palmetto via an API for parallel processing. The job data is pulled from an API using a script running on Palmetto supercomputer and sent to Matlab (SPM12) scripts to generate the secondary dataset. First, during the secondary image generation process, different 3D MRIs, as described in the subsections below, get generated along with surface data. Next, the modulated white matter (MWP1) image generated is then correlated with an IXI template image and the original image against the other images to find duplicate subjects. After correlation analysis, the next step is brain volume estimation. The data from brain volume estimation are gray matter volume (STS and OFC volumes described in detail in brain volume generation chapter). The script then updates the estimated brain volumes for a subject
in the database in the consortium using update API. The script then zips and uploads generated secondary datasets generated to the consortium file system using the secondary dataset upload API.

5.3 Different Types Of Secondary Images

Dyslexia Data Consortium generates different kinds of secondary datasets. In the Dyslexia Data Consortium back-end, python calls Matlab scripts that call the SPM12/CAT12 code for secondary dataset generation. These secondary datasets result from a series of transformations described in the sections below.

5.4 Native Space Gray/White Matter

Tissue segmentation is a process of separating an image into different tissue class segments. It is an essential process for the study of the activities of neurons. For example, the tissue classes can be lung tissue class, liver tissue class or muscle tissue class, etc. In the case of an unhealthy person, it can also be segments of tumors or inflammation[35]. Human brain tissues consist of five tissue types, blood vessel, non-brain, cerebrospinal fluid, white matter, and gray matter, [7]. Gray Matter is an integral part of the Central Nervous System (CNS) and is composed of neurons and supporting glia. It is the location for the cell bodies of neurons and gray matter and is oversees functions such as muscle control. It coordinates activities such as hearing, seeing and also controls speech recognition, decision, self-control and emotions. White matter is another essential component of the Central Nervous System. It comprises of glial cells and the axons. White matter is present around the deep parts of the human brain and the spinal cord. The fatty substance (myelin) surrounding the nerve fibers (axon) gives the white color to the white matter. The axons in white matter allow neurons to communicate with other neurons and motor and sensory information to leave and enter the brain.

Native space refers to the region of an image in original space, i.e., not transformed by some registration or normalization function. Native Space analysis gives an alternative to spatial normalization and smoothing steps in studies to analyze the white matter and the gray matter[3]. Using SPM, we can perform tissue segmentation to get the gray matter and the white matter images in native space[5]. Dyslexia Data Consortium uses SPM’s tissue segmentation to find gray matter and white matter images.
5.5 Forward/Inverse Warped Images

Image warping is a process of manipulating an image such that structures or shapes in an image are distorted to fit similar but different-sized shapes in a template image. There are several ways to warp an image. It can be translation, rotation, changing the aspect, affine transformation, or changing the perspective\(^1\). Parametric or global warping also includes operations such as scaling and rotation.

In the forward warping deformation, we transform a native space image into a normalized space. Given a coordinate in the source image \((x, y)\) using a transformation function \(T\), we can move the image from \((x, y)\) to \((x', y')\). Forward warping transformation can be represented by: \((x', y') = T(x, y)\), where \(T\) can be any one or combination of transformation operations listed previously. In inverse warping, we transform a distorted image into native space.

5.6 Modulated and Normalized Grey/White Matter

CAT12 Voxel-based morphometry is an approach to examine differences in brain structure. In VBM, every brain is spatially normalized to a template image to compare common brain locations across subjects\(^2\). Spatial normalization helps deform the brain structure. This deformation makes it possible for a location in the brain image of a subject map to the exact location in the brain image of a different subject. Spatial normalization allows local areas to stretch and compress and creates a deformation field. This deformation field can map the distance a voxel in an image must move to get mapped at some location in the template image. Modulation in VBM helps scale voxels based on their deformation. The scaling is calculated based on the compression or stretching applied during deformation. Calculating a Jacobian determinant of the deformed field is required during modulation.

5.7 Modified Native Grey/White Matter

Medical imaging techniques provide insights related to anatomical aspects of the human body. However, different imaging studies can exhibit different spatial characteristics. This differ-

---

\(^1\)http://graphics.cs.cmu.edu/courses/15-463/2006/fall/www/Lectures/warping.pdf

ence in spatial features adds a layer of complexity for clinicians in assimilating the information. As such, a process called 'Image Registration' brings images into spatial correlation[26]. This process allows physicians to compare two datasets of the same section or datasets of different subjects. The image registration process consists of 3 main steps: 1) spatial transformation, 2) Registration basis for characterizing types of features, and 3) Optimization. The performance of Voxel-Based Morphometry is dependent on the performance of registration. DARTEL stands for 'Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra'. DARTEL normalization is a commonly used spatial normalization technique\(^3\). Direct Cosine Transform (DCT) normalization offers very little flexibility, and inverse transformations not well defined. DARTEL normalization comes as an alternative to DCT and offers careful modeling and regularization. One of the common uses of DARTEL is inter-subject registration of brain images\(^4\). It also improves normalization for VBM. DARTEL normalization can use Modified native gray matter and white matter images produced by SPM 12.

5.8 Normalized Bias Field Corrected Image

A bias field is a low-frequency signal that makes MRI images brighter than the others due to the magnetic field distortion. The distortion of the magnetic field is due to varied head/body sizes, and the proximity of the tissue to the coil can influencing the contrast in the image. These corruptions happen more frequently with old MRI machines and are not desirable. In addition, they can affect the segmentation algorithms to produce incorrect results making the image corrupted. To remove this from the datasets, a technique called bias field correction is used [1]. The Dyslexia Data Consortium uses the SPM12 bias field correction technique, a part of the segmentation pipeline.

5.9 Surface Data

Researchers can use CAT12 surface-based analysis for morphometric feature extraction from the geometric models of the cortical surface of the human brain[19]. The input to the surface-based analysis is a T1-weighted (or T2-weighted or a combination) image. Extraction of the cortex is the first step in the surface-based analysis. The cortex forms the external layer of the human brain,

\(^3https://www.fil.ion.ucl.ac.uk/spm/course/slides10-vancouver/09_Morphometry.pdf\)

\(^4https://miplab.epfl.ch/BrainHack/Practice/Utilities/spm12/man/dartelguide/dartelguide.tex\)
and the model is called a surface model. A network of triangles with X, Y, Z coordinates includes a surface model. Each triangle is called a face, and the point of intersection of the corners is a vertex.

The algorithm extracts the cortical thickness from these coordinates. Spherical thicknesses such as left central thickness left spherical thickness, etc., are generated in the system for all the uploaded MRI images. All of these are available for users to download.
Chapter 6

Parallel Data Processing In Palmetto

Dyslexia Data Consortium uses a ubuntu machine to host the web application. In this server, processing a single nii image takes around 3 hours, and it consumes around 100% of the CPU capacity. Due to this, processing a large dataset would take a very long time. Since users want datasets to be available as fast as possible, parallel processing is essential to expedite the generation of secondary datasets. Dyslexia Data Consortium uses Palmetto Super Computing to perform the parallel computation of secondary images.

6.1 Palmetto Super Computer

Palmetto is a supercomputer located at Clemson University\(^1\). Many students and researchers utilize this supercomputer to conduct researches that require high computing power. Palmetto consists of 2021 compute nodes totaling about 23072 CPU cores. Of all the nodes, 595 nodes have 2 Nvidia Tesla GPU. In total there are 1194 GPU in the cluster. The official palmetto documentation\(^2\) has more details about the supercomputer. Dyslexia Data Consortium uses palmetto supercomputing resources for parallel generation of the secondary datasets, which would otherwise take days for processing. We develop different APIs on the server end to assist with the parallel processing of

\(^1\)https://www.palmetto.clemson.edu/palmetto/
\(^2\)https://www.palmetto.clemson.edu/palmetto/about/
secondary datasets. The description of the APIs is below.

6.2 API

Dyslexia Data Consortium has different REST APIs to facilitate the parallel processing of MRI images for secondary datasets generation. They provide information that can be helpful for parallel processing at palmetto. These APIs provide the functionality to download the image, upload the secondary image generated, update the processing status, etc. They have authentication at several layers to increase the security of the data. The API was developed using Python and Django REST framework.

6.2.1 Jobs API

The request type for the jobs API is GET. Users need to send the authentication header for authentication while sending a GET request to this API. Once authenticated, the API returns job details (id of the image uploaded by user, image download URL, userid of the user uploading the image, etc.). Then, a job processor uses the information for further processing.

6.2.2 Upload API

The request type for the upload API is POST. Consumer of this API needs to send a zip file containing all the secondary images generated and the job id and authentication parameters in a post request in this URL. Once authentication is complete, the Dyslexia Data Consortium processes the secondary datasets generated and stores them for further usage.

6.2.3 Download API

The request type for the download API is GET. Consumer of this API needs to send "jobid" in the API endpoint and authentication parameter in the header. Once a successful GET request is sent, dataset download starts.
6.2.4 Update API

This API serves as an endpoint to update different parameters such as correlation coefficients calculated among the dataset and brain volumes calculated. Authenticated users can only access this API. It consists of 4 endpoints, the first to update the correlation coefficients calculated. The second updates the surface data calculated. The third endpoint updates the predictions generated. The last API updates the brain volume estimated. The endpoint to update the correlation coefficient accepts POST request only, whereas the endpoint to trigger the surface data load accepts GET request only.

6.2.5 Status API

During the secondary image generation process, there are different stages. The purpose of this API is to save all the steps of data processing. This API endpoint takes a job id and status parameter, and status message and saves it to the related job.

6.3 API Format

<table>
<thead>
<tr>
<th>API</th>
<th>jobs</th>
<th>upload/download</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>/api/jobs</td>
<td>/dataset/updown/&lt;jobid&gt;</td>
<td>/dataset/status/&lt;jobid&gt;</td>
</tr>
<tr>
<td>GET</td>
<td>Returns JSON containing job details</td>
<td>Downloads 3D MRI dataset</td>
<td>Returns JSON containing processing status of job.</td>
</tr>
</tbody>
</table>

Table 6.1: API format for jobs, upload, download and status API

Table 6.1 shows the API format for jobs, upload, download, and status API. Jobs API accepts GET requests and returns JSON containing details of a job to be processed. Upload/Download API accepts both the GET and POST request. The GET request downloads the dataset, whereas it uploads the dataset to the server on the POST request. GET and POST both require the parameter "jobid" since we either download a specific dataset or update the dataset. The status API accepts both GET and POST requests. On GET request, it returns the JSON containing processing status details of the job. In contrast, it takes the status code and status message on POST and updates it.
to the corresponding job mentioned by parameter "jobid" in the API endpoint.

<table>
<thead>
<tr>
<th>API endpoint</th>
<th>correlation</th>
<th>surfacedata</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
<td>/correlation</td>
<td>/dataset/processxml/&lt;jobid&gt;</td>
</tr>
<tr>
<td>GET</td>
<td>–</td>
<td>Triggers secondary dataset parameters load into database from xml file. Requires job id in the url.</td>
</tr>
<tr>
<td>POST</td>
<td>Updates correlation coefficient between datasets. Requires dataset ids and correlation coefficients in POST data.</td>
<td>Uploads zip containing secondary dataset generated. Requires zip file 'dataset' in the post message.</td>
</tr>
</tbody>
</table>

Table 6.2: API format for correlation and secondary dataset parameters update api

Table 6.2 shows the format for update API. It consists of multiple endpoints. The first endpoint updates the correlation coefficient among datasets. This endpoint accepts POST request only and requires users to send a dataset id for which the correlation coefficient is to be updated in the database. The second endpoint triggers secondary dataset parameters update in the server. These dataset parameters are related to the subject’s surface data and are uploaded to the server as an XML file using the upload API after secondary dataset generation completes. This API accepts GET request only and requires "jobid" as a parameter in the endpoint.

### 6.4 Parallel Processing Architecture

Figure 6.1 shows a step-wise representation of job processing in Palmetto. At first jobs API of Dyslexia Data Consortium is requested by job processor setup on Palmetto supercomputer. Then the consortium returns details about the job. Once the job processor gets jobs, it starts processing the jobs in parallel, as shown in the figure. After the secondary dataset generation is complete, the job processor calculates brain volume estimation and the correlation coefficient. Then the job processor updates the values of the volumes and the coefficient calculated to the consortium. During different steps shown in the figure, the job processor sends an update of the status to the palmetto supercomputer. When all the stages of dataset generation are complete, the results are uploaded back to the Dyslexia Data Consortium. Researchers can later download all the datasets using the download page in the consortium. Introducing Palmetto for parallel processing is very helpful. It
Figure 6.1: Parallel processing using Palmetto step wise helps us speed up the secondary dataset generation process so that the researcher can access the secondary dataset and other parameters related to the data as fast as possible. On our analysis, the speed has been very significant, up to 36x speedup in the processing.

6.5 Performance Analysis

Many operations on Dyslexia Data Consortium are CPU as well as memory intensive. During a secondary image generation process using the CAT12 toolbox, it loads a nii image into RAM. All the operations are using matrix operation internally. Matrix operation of a 3D image requires a considerable amount of memory. Typically during the secondary dataset generation process from a nii image, it occupies around 3 Gb of RAM and consumes almost 100% of CPU for 3 hours. To overcome this, we used to use parallel processing technique on Palmetto. An API on the consortium allows users to download the dataset and process datasets in parallel using Palmetto.

There are two Intel Xeon CLX 6258R processors in the node which we purchased on the palmetto supercomputer. These processors have a total of 56 CPUs. There are also 2 NVIDIA A100 GPUs, and this node has 384 GB DDR4 RAM. So we can theoretically process at least 56 jobs (i.e., 56 datasets) simultaneously using the 56 CPUs. However, if we process such a large number of datasets simultaneously, we found that the consortium server, which does not have powerful hardware, keeps crashing while serving the upload request. Over a test conducted using 40 datasets, we found that processing each dataset takes an average of 1 hour in the Palmetto supercomputer.
To avoid the consortium server from getting overloaded with upload requests, we introduced a 5 minutes delay between the start of 2 simultaneous data processing jobs. As such, by introducing the delay, we limit the number of jobs that can be processed in an hour in parallel. This gives the consortium server flexibility to handle upload requests without crashing. The limitation with this approach is that we will be processing a maximum of 12 datasets only in an hour. So all together, in a day, the palmetto can process 288 datasets. Even then, this is 36 times higher than processing datasets with just a single server.

Other than that, deep learning models to detect brain tissue and facial features are also memory-intensive operations. These models are trained convolutional neural networks and are very big, almost 4 GB in size. Keeping this in memory and making predictions would be costly and makes the limited resources constantly occupied. Hence, it has been removed from the regular workflow and provided as a separate feature.

<table>
<thead>
<tr>
<th>Time</th>
<th>1 hour</th>
<th>6 hours</th>
<th>24 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Server</td>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Palmetto</td>
<td>12</td>
<td>72</td>
<td>288</td>
</tr>
</tbody>
</table>

Table 6.3: Comparision of number of datasets that can be processed in Palmetto V Single Server

Figure 6.2: Performance Analysis Palmetto V Consortium

Table 6.3 and figure 6.2 show a comparison of secondary image generation capabilities in palmetto vs. a single server. The single-server configuration takes around 3 hours to generate secondary datasets from a single dataset. In 6 hours, it can process just two images, and in 24 hours, it can process a maximum of 8 images. In a palmetto supercomputer, however, processing a single dataset takes around 1 hour on average, and with our configuration details explained earlier, we can process almost 12 datasets in an hour. So, in 6 hours, we can process 72 images, and in 24
hours, we can process 288 datasets, giving a speedup of almost 36 times.
Chapter 7

Data Quality Checks

Data sharing is essential in research and collaboration. Dyslexia Data Consortium intends to promote open access to datasets. Researchers can share different datasets related to various research on dyslexia data consortium. Data sharing facilitates research; however, data privacy and security concerns arise, mainly because participant or patient data is stored. These information should not reveal the identity of users involved in research. As such, before storing the subject’s data, they need to be de-identified. Tools such as SPM and CAT12 have limited use when it comes to the de-identification of datasets. Tools like DeId toolbox [39] uses FSL BET [38] to remove the facial features and allows users to review deidentified images at different orientations. A DeId toolbox is an effective tool for removing facial features from 3D MRI datasets and identifying features from behavior and other datasets. However, in the process of eliminating facial identification features, it can also remove brain tissues, rendering the dataset to be useless. Hence, it is of utmost priority that the dyslexia data consortium ensure the users’ privacy in the dataset and maintain the quality of the uploaded images. To ensure datasets uploaded do not have voxels representing facial features and at the same time voxels representing the brain do not get removed, a prediction system is developed at the bioinformatics lab of the Clemson university and integrated it into the consortium. This prediction system is a convolutional neural network that predicts the presence of facial features and brain voxel tissues in the uploaded images. The process that ensures the removal of facial features and brain tissue preservation is the Data Quality Check. It is an integral part of data sharing in the consortium. We can summarize the data quality check process as shown in figure 7.1.

Figure 7.1 illustrates three main steps in the data quality process. First, a user uploads a
3D MRI image. Second, after a 3D MRI dataset is uploaded, the system generates two predictions. The first one is a prediction to indicate if a face is present or not and suggest if brain tissue is present in the image uploaded. Finally, we develop, train, and integrate two convolutional neural networks in the Dyslexia Data Consortium to predict if brain tissue is present or not. Details of the data quality page, UI are discussed in the results section.

### 7.1 User Interface For Data Quality

Figure 7.2 shows the data quality check page. From this page, users can monitor the quality of the dataset uploaded. On this page, primarily four options are available. First, users can see the probability of face presence. The datasets need to be de-identified to protect the privacy of users. A probability of "0" in this column indicates that facial features are absent, whereas a "1" means that facial features are present. Second, users can see the probability of brain tissue removed. If the likelihood is a "1", then brain tissue is removed, which is not favorable, whereas a possibility of "0" means that brain tissue is present in the dataset. These probabilities are the result of the deep learning models, discussed in sections below, and are a value between 1 and 0 rounded using a threshold. Third, users can view a series of 2D images of the 3D image uploaded by clicking the link.
in the dataset name. Finally, users can delete the dataset if the dataset has brain features removed or facial features present.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Probability tissue was cropped</th>
<th>Probability face was present</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>rt</td>
<td>0</td>
<td>0</td>
<td>Ignore</td>
</tr>
<tr>
<td>MN152</td>
<td>0</td>
<td>0</td>
<td>Ignore</td>
</tr>
<tr>
<td>next_001</td>
<td>0</td>
<td>0</td>
<td>Ignore</td>
</tr>
<tr>
<td>sub-1 (anatomical)</td>
<td>0</td>
<td>0</td>
<td>Ignore</td>
</tr>
</tbody>
</table>

Figure 7.2: Data Quality Check UI

### 7.2 CNN Model Training

We trained two CNN models to predict the presence of brain tissue and facial features. The training dataset consists of images processed from IXI T1 images. First step in the dataset generation for training is the removal of facial features from the MRI images using different de-identification tools. For the de-identification of the image, we set a different values in the Brain Extraction Tool (BET) and Brain Surface Extractor (BSE) tool and run it to collect extracted datasets from each tool. Once facial feature removal is complete, we label the training dataset as having facial features or not and having a brain image not. To label the dataset we created a web page where we presented the 2D projections of the 3D MRI image from all sides to the users. Users can observe the images and then decide the presence of brain tissues and facial features. Users can then click on the "yes" or "no" button to vote about their decision. The decision that gets the maximum number of votes is the final label of the dataset. Dr. Mark Eckert and Dr. Kenneth Vaden from the Medical University of South Carolina, Dr. Zijun Wang, Akash Indani, Roshan Bhandari and Foram Joshi from Clemson University were involved in the voting process.

Figure 7.3 shows architecture of the deep learning model training process. The first layer in the CNN is a convolution layer. Including input, there are 5 convolution and max Pooling layer. We are using 2D image here so the convolution is 2D convolution. This layer consists of a stride and
Figure 7.3: CNN Model Training Process

A filter. When an input image is multiplied by a filter using a stride, it gives us a feature map. The activation layer in the convolution layer consists of a non-linear activation function. The output of the first layer, i.e., the feature map, is fed to the activation layer. Another layer is a pooling layer. A pooling layer reduces the size of an image. Once the size of the image decreases to a required format, it is flattened to a single column vector and then passed to the dense layers. Since the classification task for the Dyslexia Data Consortium requires predicting if the image has brain tissue present or not and facial tissue removed or not, we use a sigmoid layer to get the binary classification.

The model integrated into the dyslexia data consortium uses a RELU (Rectified Linear Unit) activation layer and a max-pooling layer with a (2,2) filter. The training process consisted of 1651 datasets, and the testing process consisted of 413 datasets. We used 32 for the batch size, Adam optimizer, and a learning rate of 0.01. After the training process and testing process, the facial feature prediction model had a test accuracy of 94.43% and the brain tissue prediction model
had an accuracy of 98%.

7.3 CNN Model Integration

The trained model is then exported as a file and saved in the file system. In palmetto, a python script would constantly poll the Jobs API and download the dataset. Then the script loads the model in memory. It also converts the 3D dataset into 2D dataset and gets the predictions for the 2D dataset. The script would then update the predictions back to the dyslexia data consortium by sending a POST request to the predictions API. Figure 7.4 shows different steps involved in the integration of the prediction models.

**Figure 7.4: Prediction Workflow**

7.4 Facial Tissue Present Prediction:

The images must be de-identified before uploading them to the Dyslexia Data Consortium. De-Identification ensures the privacy protection of the subjects involved in the research. The face removed model ensures that facial features such as cheeks, chin, ears, eyes, etc., are not present
in the image uploaded. Any image that does not meet the standard should get removed from the website. To assist users in ensuring data quality, we provide a tool where all the data sets and images uploaded by users go through a trained CNN model. The deep learning model takes 3D MRI images and then predicts 0 and 1. "0" means that the uploaded MRI image of a subject does not have a face, whereas "1" means that the subject’s image has facial features. Using these predictions, users should make sure images uploaded by them have facial features removed.

7.5 Brain Tissue Absent Prediction

MRI images need to be de-identified properly. During the deidentification process, loss of brain tissue may occur. Loss of brain tissue means that the MRI data uploaded to the consortium is useless. Removal of 3D MRI images with brain tissue removed helps increase the usability of the data uploaded on the Dyslexia Data Consortium. Dyslexia Data Consortium comes up with a trained deep learning model that predicts if the MRI image uploaded has brain tissue removed or not. For the 3D MRI input image, the model outputs a score of "0" or "1". A "0" means that the image does not have brain tissue removed, whereas a "1" means that the uploaded image has brain tissue removed.

7.6 2D Image visualization

MRI images uploaded in the dyslexia data consortium have three dimensions. For every 3D MRI image uploaded to the website, users can visualize 2D images by clicking on the specific link provided on the data quality page. First, a 3D MRI image is loaded in the memory using the nibabel python library and transformed into an n-dimensional NumPy array. Then, the library projects the NumPy array into a 2-dimension space from different sides. Finally, the generated image is rendered to a user using the matplotlib library. The nibabel library uses Marching Cube Algorithm[30] to create 2D images. The 2D images can be helpful to users in deciding the quality of the data. Figure 7.5 shows a sample 2d image generated from a 3d MRI dataset uploaded to the dyslexia data consortium. The 3d to 2d image conversion is achieved using the nibabel python library.
7.7 Image Similarity Check

A critical aspect of dyslexia research and dyslexia dataset is to ensure that the subject does not repeat. To ensure that, the dyslexia data consortium offers a solution using correlation analysis among all the subject images uploaded in the dyslexia consortium. The image similarity analysis provides a metric for how similar an image is to the other images included in the study. The similarity calculation uses the CAT12 toolbox’s covariance analysis. The higher the correlation coefficient between two images, the higher is the probability that two images were shared for one subject.

Figure 7.6: Co-variance calculation for image similarity

Figure 7.5 shows the steps involved in covariance calculation. First, a Python script downloads the 3D MRI image uploaded in the consortium and adds it to a pool of images for correlation analysis. Then the script calls Matlab and CAT12 script for covariance calculation. Once covariance calculation completes, the script updates the results to the server. Users can then view the results and determine if the subjects are the same.
Chapter 8

Brain Volume Estimation

Children with dyslexia are found to have significantly smaller gray matter volume in different sections of brain [10][40][23][27]. Some research also noted structural irregularities[34]. The difference in brain volume and structure may result in an alteration of the information processing capability of the brain. As such, it is essential to study the changes in brain volumes to understand dyslexia properly.

There are different tools and techniques are used for brain volume estimation. Freesurfer[16], FMRIB Software Library[22], Computational Anatomy Toolbox (CAT12)[9] are some of them. Some of the recent techniques include using convolutional neural networks (CNN) [11] for brain volume estimation. A researcher can use Dyslexia Data Consortium for brain volume estimation. The brain volume estimation capabilities in the Dyslexia Data Consortium come up from integrating CAT12 toolbox functionalities. First, a 3D T1-weighted image is interpolated and normalized using an affine registration. Then it is denoised, corrected, and segmented. The segmentation consists of white matter, gray matter, and CSF components[11]. In the Dyslexia Data Consortium, we use the mwp1 image to find the brain volume. It calculates the gray matter volume of two regions. They are the superior temporal sulcus and orbitofrontal cortex. We are primarily concerned with this region since research has found a significant decrease in brain volumes of these regions in dyslexic patients.
8.1 Superior Temporal Sulcus (STS) Volume

Superior Temporal Sulcus (STS) is a sulcus separating the middle and superior temporal gyrus in the left hemisphere of the brain[4]. Some researchers see it as an essential region for audiovisual integration. Some researchers view it as the central region for the theory of mind. At the same time, some others believe it to be responsible for the processing of speech and faces[20]. The volume of the Gray matter in the STS region decreases significantly in persons with dyslexia[13]. Hence it is essential to study STS volume in the dyslexic patient. Researchers can use the dyslexia data consortium to perform brain volume estimation. All the uploaded images go through volume estimation automatically. At first, a Python script creates a mask for these images using the STS template MRI provided in the CAT12 toolbox. Then it uses SPM12, and spm_read_vols\(^1\) function to estimate the volume. Finally, it saves calculated volumes in the database for further analysis.

8.2 Orbito Frontal Cortex Gray Matter Volume

The orbitofrontal cortex is found at the front of the brain and lies just above the orbits[2]. Researchers associate this region of the brain with decision making[41]. An analysis conducted by Mark et al. in patients with reading disabilities shows the reduction in gray matter volume in the orbitofrontal cortex and superior temporal gyrus [15]. Dyslexia Data Consortium using the SPM12 functions, spm_vol and spm_read_vols, calculates and saves the OFC gray matter volume in the database. OFC volume estimation uses a mask of the OFC MRI template. Researchers can use the saved information for further analysis.

8.3 Brain Volume Estimation Pipeline

A nii image goes through a step of the process, as shown in figure 8.1. For the image uploaded, the system generates secondary datasets. Using the modulate and normalized white matter image, i.e., mwp1 image and spm_read_vols function, average gray matter volume for STS and OFC regions are estimated. Then calculated volumes are saved to the database.

We can summarize Figure 8.1 in the following steps:
1. User uploads 3D MRI images in the consortium.
2. Secondary images, including modulated

\(^1\)https://github.com/neurodebian/spm12
normalized gray matter image (mwp1) is generated. 3. SPM_READ_VOLS is called along with STS and OFC mask. 4. STS and OFC volumes are calculated and written to the database (e.g., variable name STSvolume, OFCvolume) for each subject. 5. These volume data should be available for download with the demographic/behavioral data.

8.4 Brain Metrics Page

Different brain metrics are essential for dyslexia research. The Dyslexia Data Consortium provides an option to view brain metrics such as gray matter volumes. Users can view the STS and OFC gray matter volume for the dataset uploaded by them in the consortium. The user interface also provides a correlation coefficient with different other datasets uploaded in the system. This coefficient can be used to find duplicate subjects in the dataset. Figure 8.2 is the user interface to view different brain metrics for the dataset uploaded by a user.
Figure 8.2: User Interface for Brain Metrics
Chapter 9

De-Identification Toolbox

Integration

9.1 De-Identification toolbox and its Features

A De-Identification toolbox is a java software developed by Xuebo Song et al. [39]. It allows researchers to share de-identified human subject data from structural imaging studies. It removes information that could be used to identify or re-identify a subject identity. That is, it is designed to reduce the risk for subject identification.

In the De-Identification toolbox, users can use the interface to select the NIFTI image (nii format) files. After the user selects the images, users can also upload demographic and behavioral files and remove identifying information such as name or age explicitly. Identity information remains even after explicit removal. For these data, randomization and generalization approaches are used[42].

Another significant function of the De-Identification toolbox is called Associative Match. The toolbox assigns a unique id to all the subjects in data and image files during this step. The associative match is a critical process to unlink the ID value to personal health information. During data preparation, users can view variables as well as remove variable names. Some variables, such as date of birth and names, are automatically removed by the system.

The NIFTI files have voxels that represent the face. The toolbox removes these voxels during the skull stripping process. DeId uses BET[38] for skull stripping. BET is a powerful tool
that allows the skull stripping of multiple images in multiple orientations at a time. Users can also view the 2D representation of each image sliced from different positions. They can also render the image directly to MRICron and check if the skull is stripping properly or not. Users can also use the montage function to inspect the image.

Finally, researchers can also use the De-Identification toolbox to edit the header files and data description and share the de-identified data using FTP and SFTP protocol. This toolbox works with all major operating systems.

9.2 Data Consortium Upload API

De-Identification toolbox allows a researcher to remove identifying information from a nifti image. Users can skull strip their dataset and share it with the other users using the FTP and SFTP share. However, the researcher can’t share the dataset with a larger community of researchers.

Dyslexia Data consortium is a platform that allows researchers to share datasets. One of its primary purposes is to enhance collaboration and data sharing for dyslexia research. Users can upload datasets directly or generate different kinds of secondary images and download datasets required for their research. Bringing together the power of the De-Identification toolbox and the dyslexia data consortium would enable researchers with more research capabilities.

Dyslexia Data consortium has an API that allows developers to upload a dataset to it. This API takes a zip file containing the de-identified 3D MRI dataset generated from the De-Identification toolbox and authentication headers. It uploads them to the Dyslexia Data Consortium as a regular upload in the web platform of the Dyslexia Data Consortium. The dataset to be uploaded needs to be sent to the upload API using an HTTP post method. The De-Identification toolbox’s data sharing UI also incorporates features to accommodate data sharing to the dyslexia data consortium. A section with essential elements is added to the De-Identification toolbox. It includes input fields to enter the consortium username and password required to send authentication parameters in the POST header. In the same section, users can add a title and description to the dataset and select the image type. Figure 9.1, shown in the results section, shows the De-Identification toolbox UI for sharing the dataset in the Dyslexia Data Consortium.
9.3 De-Identification Toolbox UI

De-Identification toolbox allows users to remove identification features such as face and skull from a 3D MRI dataset. Using the toolbox, once a skull is stripped, and facial features are removed, users are redirected to a share page. Previously, users could only share the results with other users via FTP. This limited share feature has been updated, and users can now share the datasets on the Dyslexia Data Consortium. Figure 10.8 shows the upload page of the DeId toolbox. Users need to click on the transfer result to the Dyslexia Data Consortium checkbox to share the dataset. Once that is done, users need to enter their Dyslexia Data Consortium credentials, i.e., username and password, to the corresponding field, then give a name and a short description to the dataset, select the image type and click on the continue button. The dataset will then be uploaded to the consortium.

![DeId toolbox share page updated](image.png)

Figure 9.1: DeId toolbox share page updated

9.4 Integration Workflow

Figure 9.1 shows implementation architecture (workflow) in steps. When dataset de-identification is complete, the data share page pops up in the DeID toolbox. Users can enter their username and password, enter the title, description, and image type of the dataset, and hit the upload button. The toolbox prepares a zip file containing the images uploaded by the user and creates an HTTP
Figure 9.2: DeID Integration Workflow

POST request to send the data to the consortium. The dataset finally gets processed and saved in the consortium for further processing.
Chapter 10

Conclusion and Discussion

10.1 Conclusion

Sharing clinical and experimental data can increase collaboration and help expedite scientific discovery and the development of new methods. Dyslexia Data Consortium provides a web-based platform that helps researchers to share datasets and collaborate on research. It ensures data shared and stored are compliant with standard protocols such as BIDS to ensure data privacy. Integration of DeId toolbox helps ensure datasets are appropriately deidentified before sharing. Features such as Brain Volume Estimation, Secondary dataset generation can help researchers understand dyslexia better. Finally, we also plan to release Dyslexia Data Consortium as an open-source tool allowing more features to be integrated into the platform that would support more kinds of analysis. Dyslexia Data Consortium open source is available at: https://github.com/bioinformatics-AI/dyslexia_website

10.2 Discussion

10.2.1 Testing and Evaluation

Dyslexia Data Consortium was tested extensively by our team of developers and collaborators. The group includes collaborators across different universities around the world. Dr. Mark Eckert, Dr. Kenneth Vaden, Tanner Glaze from the Medical University of South Carolina, Dr.
James Wang from the Clemson University have also tested and verified different workflow of the consortium. The testing was extensive and covered all features, i.e., data upload, data download, data quality predictions, brain volume estimation, correlation analysis, upload from DeId toolbox, secondary images generation, etc.

10.2.2 Distribution

The website will continue to be developed and tested at the bioinformatics lab at the Clemson University for some time. It will also be released as an open-source tool, and the code will be available on GitHub. Any researchers and developers would be able to contribute to the development of the consortium. The De-Identification toolbox is available on the NITRC website and the consortium to the investigators for download.

10.2.3 Future Works

The dyslexia data consortium aims to bring together researchers for collaboration on dyslexia research. Currently, the consortium has basic features required for collaboration. However, we can add advanced features such as visualization of different data, allowing researchers to interact on specific dyslexia datasets. We can also develop tools to store and share variables for studies of other developmental or neurodegenerative disorders on the system. We can also enhance the deep learning-based data quality to include behavioral datasets and other metrics to get more accurate predictions. And finally, deep learning and machine learning models to predict if any dataset uploaded is dyslexic or non-dyslexic can be developed and integrated.
Bibliography

[5] Tissue class segmentation — Introduction to MRI.


