WPI

Deep Learning Analysis on Neuroimaging Data to Distinguish Anxiety and Depression Diagnoses in Adolescents

A Major Qualifying Project
Submitted to the faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Bachelor of Science

By: Olivia Deckers
Pitipat Kongsomjit

Date: March 2024

Report Submitted to:
Dr. Benjamin Nephew, Advisor

Acknowledgements

We would like to thank our advisor, Dr. Nephew, for his guidance and expertise throughout this research project. We would also like to thank Senbao Lu and Bhaavin Jogeshwar for their technical assistance with data preprocessing and the deep learning pipeline. This project would not have been possible without everyone's support and previous project work.

Table of Contents

Abstract

Depression and anxiety disorders are two of the most diagnosed mental disorders in adolescents. These disorders can be debilitating and put significant burdens on the quality of life of the children and parents. Pathological changes in the brain's structural anatomy have been investigated using Magnetic resonance imaging (MRI). Statistical analysis, machine learning, and deep learning techniques have identified key regions of interest in the brain whose abnormalities are associated with these disorders. The Adolescent Brain Cognitive Development (ABCD) study is the largest longitudinal study of childhood brain development in the United States. Using data provided by ABCD, this study aims to differentiate between anxiety and depression diagnoses and identify key regions of interest in the brain using a 3D Convolutional Neural Network and Random Forest analysis.

Introduction

1. Anxiety and Depressive Disorders in Adolescents

Depression and anxiety disorders are two of the most diagnosed mental disorders in adolescents, with pooled prevalence rates of 8.5% and 11.6%, respectively (Racine et al., 2022). These disorders can put significant burdens on the quality of life of the individual and parents. In adolescents, depression is a critical risk factor for suicide with more than 50% of adolescent suicide victims being reported to have a depressive disorder (Petito et al, 2020). Depression can also lead to serious social and educational impairments, such as an increased rate of smoking, substance misuse, eating disorders, and obesity. In 2018, it was estimated that the average yearly family financial burden due to depression (including inpatient care, general practice and specialized outpatient care, and medication) was $16,170 per adolescent (Bodden et al, 2018).

In many children, these disorders are comorbid, making them often difficult to distinguish clinically. Additionally, because adolescents are in the developmental stage, their behavioral symptoms may not always be properly diagnosed. In recent years, particularly through the COVID-19 pandemic, prevalence of anxiety and depression have increased, emphasizing the importance of effective diagnostic and preventative measures (Racine et al., 2022). Pooled prevalence estimates of clinically elevated depression and anxiety symptoms for adolescents in 2021 were 25.2% and 20.5% (Racine et al., 2022).

2. Diagnostic Measures of Depressive Disorders

Depressive disorders are diagnosed and distinguished from general mood fluctuations by clinicians using the Diagnostic and Statistical Manual of Mental Disorders, DSM-5. Disorders are defined as a "syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning" (DSM-5, 2013). Amongst depressive disorders, many clinical subtypes exist to classify length, severity, and comorbidity of symptoms (Benazzi et al). Bipolar depression and (unipolar) depressive disorders make up the central two subtypes of depression. Bipolar depression is categorized into bipolar I disorder, bipolar II disorder, cyclothymic disorder, and bipolar disorder due to other medical or substance disorders (DSM-5, 2013). Unipolar depressive disorders are largely classified into three groups, major depressive disorder (MDD), persistent depressive disorder (PDD), and depressive disorders due to other medical conditions.

3. Depressive Disorder Subtypes

For adolescents, Major Depressive Disorder is currently the fourth leading cause of illness and disability worldwide, with an estimated prevalences between 5 and 8% (Shorey et al., 2021). MDD is characterized by persistent sadness, low motivation, hopelessness, and lower cognitive functioning. It can have severe consequences on an individual's psychological health and is highly comorbid with additional physical and mental health problems such as diabetes, heart disease, and anxiety. (Liu et al., 2022) In severe cases, individuals with MDD have increased thoughts of death, suicidal ideation, and self-harm proneness with a 25-fold higher risk of

committing suicide compared to their healthy counterparts (Zubrick et al., 2017). 30% of youth with MDD reported some form of suicidality in 2022, and over 10% reported a suicide attempt (Walter et al., 2023).

Persistent Depressive Disorder, also referred to as PDD or dysthymia, is a common, chronic depressive disorder with estimated prevalence rates between 1.6 and 8% (Forrest et al, 2021). Persistent depressive disorder expresses milder depressive symptoms but for two years or longer, making it often more disabling than major depressive disorder. Individuals with PDD oftentimes show symptoms of poor appetite, insomnia, fatigue, poor concentration, and hopelessness (Shorey et al., 2021).

4. Anxiety Disorders and Diagnostic Measures

Anxiety disorders in adolescents are diagnosed by clinicians through questionnaires, interviews, and rating scales based on criteria from the DSM-5. Many clinical subtypes of anxiety disorders exist including panic disorders, agoraphobia, separation anxiety, social anxiety, selective mutism, specific phobias, and generalized anxiety disorder (DSM-5, 2013).

5. Anxiety Disorder Subtypes

Social Anxiety Disorder is characterized by the excessive fear of being humiliated or scrutinized by social peers. By the end of adolescence, prevalence of social anxiety disorder is estimated to be around 10% (Leigh & Clark, 2018). Generalized anxiety disorder (GAD) is a common disorder in adolescents with prevalences estimates ranging from 2.9% to 4.6% (Gale et al, 2016). GAD is characterized in adolescents by excessive worrying about uncontrollable everyday events lasting for at least six months. Separation Anxiety Disorder (SAD) is characterized by developmentally excessive fear or anxiety concerning separation from attachment figures. Prevalence rates of separation anxiety disorder are estimated to affect approximately 1.6% of adolescents (Gale et al, 2016).

6. Treatment Standards for Anxiety and Depression in Adolescents

Anxiety and depression in adolescents are most effectively treated with a combination of pharmacologic and nonpharmacologic treatments (Pettitt et al., 2022). Cognitive behavioral therapy, the current evidence-based standard for adolescents with anxiety, combines therapeutic approaches that “teach the patient to identify feelings” and “cultivate mechanisms for modifying these thoughts” (Pettitt et al, 2022). Other less popular forms of nonpharmacologic treatment include social effectiveness therapy for children (SET-C) and acceptance and commitment therapy (ACT). The only approved medication for generalized anxiety treatment in adolescents is duloxetine, a serotonin-norepinephrine reuptake inhibitor (SNRI), although studies have shown that serotonin reuptake inhibitors (SSRIs) are more effective against anxiety in pediatrics. Compared to adult patients, however, the efficacy of pharmacologic treatments has not been thoroughly studied.

For mild cases of depression, psychotherapy alone is used to treat adolescents. CBT is the most popular nonpharmacologic treatment followed by interpersonal therapy (IPT) and ACT. For severe cases, antidepressants complement psychotherapy, most commonly SSRIs. Again,

efficacy and tolerability are not well studied in adolescents. A combination of other lifestyle changes including diet and exercise are also advised according to the patient's needs.

Due to high symptom variability in anxiety and depression disorders in adolescents, guidelines on personalized treatment in limited (Pettitt et al., 2022). With limited research on the adolescent population and increased challenges to obtain accurate diagnoses, determining the proper course of treatment can be challenging and time-consuming for physicians.

7. Implicated Neuroanatomical Features for Anxiety and Depression

Many neuroanatomical changes have been associated with anxiety and depression diagnoses in adolescents. The heterogeneity of anxiety and depression disorders is reflected in the variety of implicated brain regions in literature. Additionally, because they are highly comorbid, their co-occurrence is known to show unique brain patterns on MRI (Schmidt et al, 2018).

In previous studies, structural MRI demonstrated that the gray matter volumes of the left thalamus, right thalamus, and right amygdala in MDD subjects were smaller than in healthy controls. Structural and functional impairments in the right amygdala were correlated with symptoms of depression and its cognitive ability, also predicting treatment outcomes (Li et al, 2022). Additionally, "reduced hippocampal volumes and lower grey-matter volumes in cortical regions" were associated with depressive symptoms (Pagliaccio et al, 2020). Adolescents with anxiety disorders have been shown to have larger gray matter volumes in the dorsal anterior cingulate and had decreased gray matter volumes in the inferior frontal gyrus (ventrolateral prefrontal cortex), postcentral gyrus, and cuneus/precuneus (Strawn et al., 2015). Other regions shown to have a significant correlation with generalized anxiety disorder in adolescents include the left amygdala, temporal gyrus, and bilateral amygdala (Gerlach & Gloster, 2020). Depending on the severity of anxiety, however, some studies have found no significant correlation between structural brain imaging and anxiety diagnoses (Gerlach & Gloster, 2020).

In standard morphometric analyses of brain structure differences, it is common for sMRI images to be processed using software such as FreeSurfer (http://surfer.nmr.mgh.harvard.edu/). This software provides a processing stream for MRI images that includes skull stripping, labeling of regions, registration of the cortical surface with an atlas, and statistical analysis of group morphometry differences. Statistical analyses such as multiple regression between symptoms and ROIs or chi-squared are performed to determine statistical significance and implicated brain regions for anxiety and depression. These methods are often limited, however, by small sample sizes, high variability in disease severity, and an inability to make causal inferences about brain structure patterns. Additionally, many studies rely only on self-reported statistics in place of standardized psychiatric interviews. This can be challenging with adolescent patients due to an inability to articulate experiences and the developmental nature of the adolescent period. Thus, such methods would not likely be used clinically to support diagnostic efforts for children.

8. Machine Learning on MRI Images – Mental Health Disorders

Due to diagnostic limitations and increased disease prevalence, brain structure diagnosis tools have become of interest in diagnosing mental health disorders. Machine learning and deep

learning models have been designed to classify diagnosed patients from healthy controls based on magnetic resonance imaging (MRI) brain scans.

Magnetic resonance imaging is a noninvasive medical imaging test used to view detailed images inside the body. Structural MRI (sMRI) provides images of the brain's anatomical structures including gray matter, white matter, and cerebrospinal fluid, allowing for the assessment of overall brain morphology, size, and the presence of abnormalities or lesions. Functional MRI (fMRI), however, measures the small changes in blood flow that occur with brain activity. Unlike structural imaging, fMRI imaging can track brain connectivity between various brain regions. From structural images, measurements of specific brain regions can be used to provide insights into the etiology of some of the most common mental health disorders such as depression, ADHD, and anxiety. Both deep learning and traditional machine learning methods have been used to develop diagnostic classifiers for mental health disorders in the past based on both structural and functional MRI data (Li et al., 2023). For this study, only structural brain images were considered.

Although machine learning studies use a variety of brain features, feature extraction, and feature selection methods, many studies test one to five feature selection algorithms and choose the optimal N features according to their classification model (Bondi et al., 2023). Most commonly, an atlas is used to parcellate neuroimages into distinct brain regions, and a program would then calculate the volume of each region amongst other statistics such as thickness and surface area as a feature extraction step. A region-based feature selection process would then be applied, typically using least absolute shrinkage and selection operator (LASSO) as seen with (Gai et al., 2022) and (Li et al., 2023). The final step is to feed the selected features into the classifier of choice.

The most common machine learning classifier for mental health diagnosis is support vector machines (SVMs). In 2017, Hilbert et al. used SVMs to distinguish health controls from MDD and generalized anxiety disorder, achieving an accuracy of 68.05% for disorder classification. In 2022, sMRI images were used to distinguish MDD from bipolar disorder (BD) using an SVM model for multimodal and single modality classifiers. Researchers were able to achieve 85.9% accuracy with single-modality data (Liu et al, 2022). When distinguishing anxiety disorders from health controls in adolescents, SVMs were able to reach an accuracy of 69.17% (Zhang et al, 2022). Other popular machine learning models include random forest, linear discriminant analysis, and XGBoost. Depression trajectories were predicted by Xiang et al. with fMRI data using K-Nearest Neighbor, Random Forest, Gradient Boosting Machine, and Extreme Gradient Boosting models to achieve an accuracy of 87% (Xiang et al., 2022). In these studies, however, several limitations were discussed, particularly the small sample size and heterogeneity of disease severity. Heterogeneous illness severity, medication, duration of illness, and comorbidities could have impacted the accuracy of the classification models. Additionally, machine learning methods applied in this domain are based on features computed through region parcellation via an atlas. Even highly detailed atlases cannot utilize the spatial resolution of the 3D brain images that are smaller than 1mm (about 0.04 in) voxels and can be limiting when fitting a brain to the atlas. Many studies implement 5 or 10-fold cross-validation in the training stage to optimize model parameters (Liu et al, 2022). The final evaluation is characterized

largely by area under the curve (AUC), accuracy, balanced accuracy, specificity, and F1 score. Given a lack of standardization in atlases and normalization methods, comparison between models and methods is difficult.

9. Deep Learning on MRI Images – Mental Health Disorders

Notably in recent years, deep learning 3D convolutional neural networks (CNN) saw a rise in popularity due to their state-of-the-art performance in vision tasks. This popularity has carried over to neuroimaging domains as well. In 2015, 431 papers were published in IEEE-Xplore containing "medical imaging" and "deep learning" keywords. This number doubled the next year, reaching a peak interest of 949 papers (Bernal et al., 2019).

The primary reason for the success of deep learning and CNNs is their 'built-in' feature engineering. Traditional machine learning focusses on using domain knowledge and intuition from experts alongside other feature engineering methods to provide low-dimensional input vectors to models. However, the rise of representation learning has shown that deep neural networks can learn better representations capable of encoding information better within a high dimensional vector. Within the context of neuroimaging, the CNN takes the raw, non-feature extracted images as inputs and over iterations learn a way of representing neuroimages as a high dimensional vector such that all essential information is captured. This means region parcellation via atlas is no longer required as the network is learning features on its own.

Evaluation metrics for deep learning are split between variations of cross-validation and held-out tests, but studies employing variations of cross-validation tended to report higher performance compared to held-out test studies, which Zhang-James et al., 2023 suggests may be an overestimation of the model's performance. Cross-validation (CV) methods are often used for model selection or hyperparameter tuning, i.e.: a model configuration is setup, error for that configuration is estimated via some CV method, then it is repeated for all configurations. The model which minimizes the computed CV error is then selected as the final model. However, a common problem with using the computed CV error as an estimation of the model's true performance is data leakage because the model parameters are being tuned based on information in the validation sets. For classification tasks such as disease diagnoses, accuracy, precision, recall, and F1-score are most used to evaluate a neural network on the model's test data.

In 2022, Hong et al. used a 3D CNN against structural MRI images to distinguish adults with major depressive disorder from health controls with 85% accuracy. More recently, a 3D CNN classified patients with schizophrenia from health controls with 92.1% accuracy (Zhang et al., 2023). Here, both whole head scans and whole brain scans were tested for accuracy. T1-weighted whole brain scans proved the most effective. Both studies proved a 3D CNN deep learning model to be more accurate than traditional region-specific machine learning models. There is limited work using 3D CNNs for anxiety diagnoses and research specific to adolescents.

10. Brain Structure vs. Symptom-Based Diagnoses

Although symptom-based diagnoses of anxiety and depression disorders can be effective, it is often a lengthy and somewhat unreliable process in adolescents. Particularly, clinicians often have concerns that young people cannot articulate their symptoms and tend to rely on parent questionnaires and adult interviews (Garber and Weersing, 2010). Also, there is high comorbidity between anxiety and depression in adolescents with overlapping symptoms such as irregular sleep schedules and lower school performance. These factors make it particularly difficult to accurately diagnose anxiety and depression in adolescents, although proper diagnoses must bevery important that proper diagnoses are made. The prognoses for comorbid anxiety and depression are worse than either disease along with longer disease duration, more severe symptoms, and increased risk of recurrence (Garber and Weersing, 2010).

Brain structure analyses from MRI images have been proposed as a complementary tool to be used in addition to standard statistical analysis procedures for classifying anxiety and depression and determining regions of interest. Although machine learning and deep learning models are trained using supervised learning, it is believed that these models can be used to help identify regions of interest and brain abnormalities that distinguish various mental health conditions. Additionally, these tools may be effective in diagnosing disease severity and predicting disease progression or treatment response (Liu et al., 2022).

The purpose of this research, using structural, T1-weighted MRI data collected by the Adolescent Brain Cognitive Development (ABCD) study, is to distinguish between depression, anxiety and co-occurring diagnoses cases in adolescents using an approach based on 3D deep convolutional neural networks. We believe that most studies have been conducted on datasets sampled from an adult population, and many are hindered by small datasets resulting in overfitting of the model. Few studies distinguishing anxiety and depression diagnoses from MRI images have achieved an accuracy high enough to be considered clinically significant, and early diagnosis of mental health disorders is crucial to patient care and outcomes.

Methodology and Materials

1. Adolescent Brain Cognitive Development Study

The data used in this study was obtained from the Adolescent Brain Cognitive Development (ABCD) Data Release 5.1 (*ABCD Release Notes 5.1*.). The Adolescent Brain Cognitive Development (ABCD) study is the largest long-term study of brain development and child health in the United States. Funded by the National Institutes of Health, the ABCD study has involved 11,880 children ages 9-10. From adolescence to young adulthood, researchers track participants' biological and behavioral development to inform and promote the health, well-being, and success of children. https://nda.nih.gov/abcd). The ABCD study is the largest long-term study of brain development and child health in the United States. Funded by the National Institutes of Health, the ABCD study has involved 11,880 children ages 9-10. From adolescence to young adulthood, researchers track participants' biological and behavioral development to inform and promote the health, well-being, and success of children.

The ABCD protocol is a comprehensive set of physical, cognitive, social, emotional, environmental, behavioral, and academic assessments, as well as multimodal neuroimaging and biospecimen collection for hormonal, genetic, epigenetic, environmental exposure, and substance use analysis. These assessments are anticipated to be done annually (non-imaging) or biannually (imaging and bioassays) for 10 years. Adolescent participants also undergo a brief, mid-year phone interview (ABCD Study, 2024).

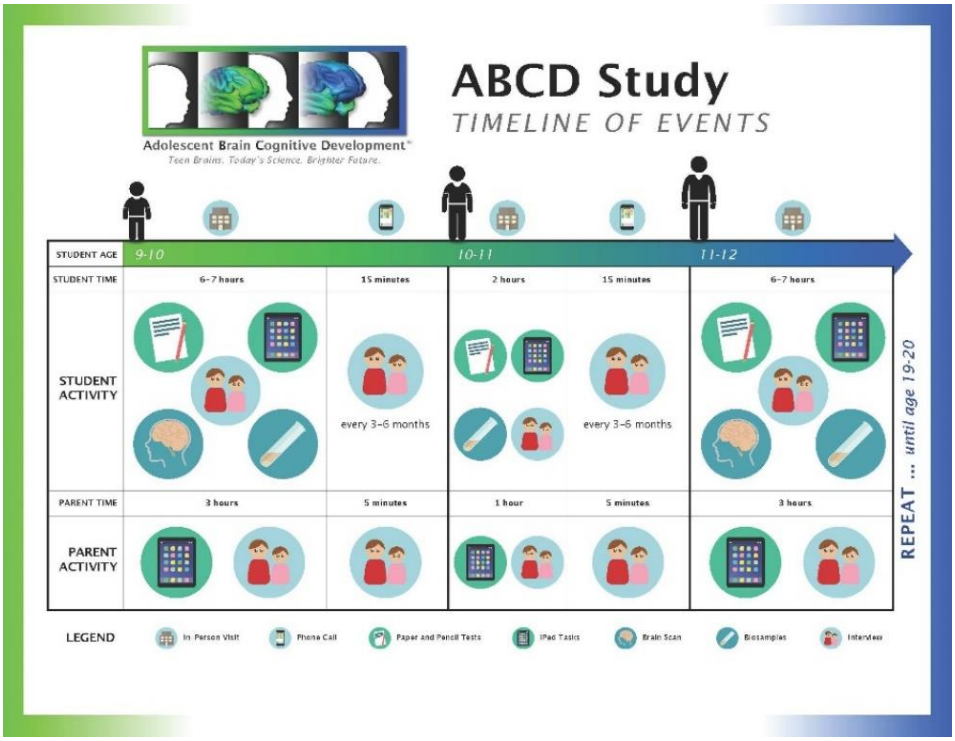

Figure 1. ACBD Study Timeline of events. Taken from https://abcdstudy.org/scientists/protocols/.

2. Overall Study Design

The study proposed three central binary classification problems: healthy controls vs. depressive disorder, health controls vs. anxiety disorder, and depressive disorder vs. anxiety disorder. ABCD provides diagnosis data from a baseline assessment and 1-year, 2-year, and 4-year follow-up assessments. Using this timeline, we were able to distinguish a third category of participants which we considered to be "predisposed." These data included individuals that were not currently diagnosed with a depressive or anxiety disorder but later developed a disorder in a follow-up assessment. [OBJ] Due to concern that a previous diagnosis could influence model accuracy, predisposed patients were removed from the study. Healthy controls consisted of subjects that were never diagnosed with either an anxiety or depressive disorder.

Study methodology consisted of four parts (1) Data Preprocessing and Preparation, (2) Data Augmentation, (3) Deep Learning Pipeline, and (4) Evaluation, Interpretation, and Discussion. Imaging data preprocessing was completed using the Clinica software (https://www.clinica.run/) as implemented in step A of the T1-volume pipeline. Preprocessing on raw T1-weighted sMRI images is important for good classification model performance, to standardize images, and to help with research reproducibility.

All preprocessed images applied bias correction, spatial normalization, and skull-stripping. We then augmented the imaging data before each deep learning iteration by applying random Gaussian blurring, and random cropping. We also experimented with cutout augmentation using voxels of size 96 x 96 x 96. The augmented images were then passed onto the deep learning pipeline, consisting of a series of CNN layers leading to a final embedding vector. This embedding vector was then passed onto the fully connected layers which eventually outputs the diagnosis prediction. To interpret our results, we made use of saliency maps which helped us understand what brain regions were used most heavily in the model prediction. Additional evaluation metrics such as balanced accuracy and AUC were analyzed to assess the model.

3. Data Preprocessing

All scans obtained from ABCD were formatted in the BIDS standard, allowing for easy compatibility with the Clinica preprocessing software. Bias correction and spatial normalization to the Montreal Neurological Institute (MNI) template was performed using step A of the T1-volume pipeline in Clinica. The preprocessed images were 121 × 145 × 121 voxels, with a voxel size of 1.5 × 1.5 × 1.5 mm^3.

Patient diagnosis data was retrieved and prepared from ABCD Tabulated Release Data 5.1 under the mh_y_ksads_ss table (https://nda.nih.gov/study.html?id=2313). The ABCD study provides symptom and diagnosis data for study participants using K-SADS scores. K-SADS is a comprehensive, semi-structured interview tool used to assess and diagnose a variety of psychiatric disorders found in children. (Note: K-SADS 2.0 was administered beginning at the 3-year follow up interview.) For this study, Major Depressive Disorder, Persistent Depressive Disorder, Other Specified Depressive Disorder, and Unspecified Depressive Disorder subgroups were used to determine depressive disorder diagnoses. Generalized Anxiety Disorder, Social

Anxiety Disorder, and Separation Anxiety Disorder subgroups were used to determine anxiety disorder diagnoses. ABCD tabulated data was filtered to include only rows that had corresponding images. Participants were also filtered to address class imbalance so that there were roughly twice as many control subjects as non-controls.

K-SADS Field Value	Clinical Interpretation
0	Negative Symptom/Diagnosis
1	Positive Symptom/Diagnosis
555	Not administered
888	Deliberately not asked

Table 1: K-SADS Field Values in mh_y_ksads_ss Table.

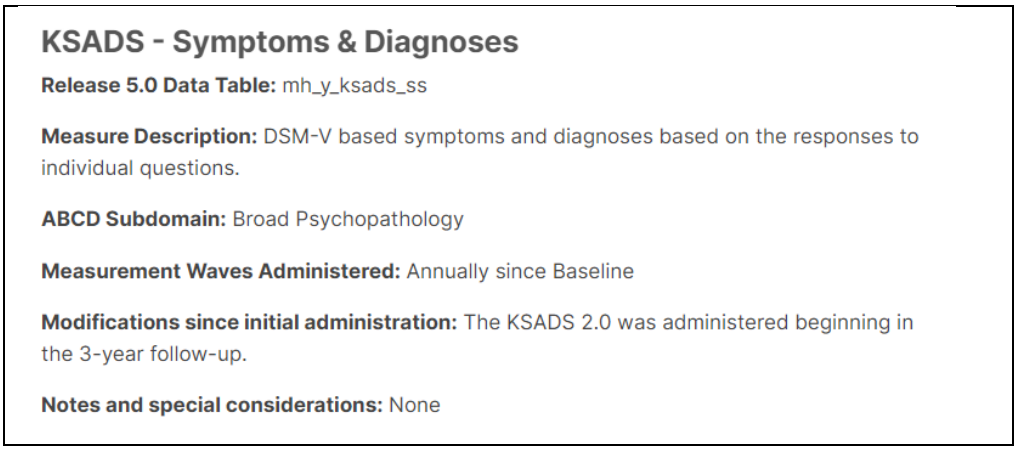

Figure 2: ABCD Study Description of K-SADS Data

4. Data Augmentation

To address the low number of training examples, we are applying random cropping and gaussian blurring methods following its success described in (Liu et al., 2022). We believe the success of random cropping owes to three of its benefits in context of the given task.

(1) Increasing the number of training examples while decreasing the number of input dimensions tends to alleviate overfitting, which is especially important given that we are dealing with very few training examples and large, 3D images. (2) Because different MRI machines have different image resolutions, but the CNN architecture expects a fixed-size input either cropping, padding, or some other processing step is necessary to fit the image in the model. Random cropping combined with gaussian blurs in effect allows the model to learn a generalizable representation of images of different resolutions. (3) Because at any given time the model will only have a cropped section of the image to make a prediction from, this forced the model to make the most out of the available features without relying its prediction on the existence of a specific feature, which should reduce overfitting.

5. Deep Learning Pipeline

A 3D CNN, composed of convolutional layers, instance normalization, ReLUs and max-pooling layers, was designed to perform classification of depressive disorders, anxiety disorders, and healthy control cases. Non-neuroimage clinical information about the subjects are encoded into a vector via a separate network; the output of this network is summed elementwise with the embedding vector outputted by the final convolution block and passed onto the fully connected layer. Classification is done using a SoftMax classification head.

Transfer learning is used to improve performance, with the base model being the one developed in (Liu et al., 2022). Most notably, this model achieved SOTA performance over other CNNs for image classification in neuroimaging—likely due to the combination of utilizing instance normalization over batch normalization, small kernel sizes, and a wide network architecture that minimizes the layers of the model.

The model was trained using stochastic gradient descent with momentum of 0.9 to minimize a cross-entropy loss function. We used a batch size of 16 due to computational limitations. We used a learning rate of 0.003 with 200 epochs of training chosen by trial based on validation balanced accuracy performance of the best performing epoch. During training, the model with the lowest validation loss was selected and saved for further interpretation

6. Deep Learning Evaluation metrics/Model Interpretation

Based on findings from (Zhang-James et al., 2023), we believe the hold-out test set method to offer the most accurate evaluation of the model's performance on unseen data. AUC curves were computed per class by running the best performing model on the hold-out set. This metric indicates the relationship between the true positive rate and false positive rate when the classification threshold varies (Liu et al, 2022). AUC curves were created to distinguish between all three binary problems proposed: healthy controls vs. depressive disorder, health controls vs. anxiety disorders, and depressive disorders vs. anxiety disorders. We also calculated two types of averages, micro- and macro-average denoted as Micro-AUC and Macro-AUC respectively according to Liu et al. The micro average treats the data as an aggregated result and consists of the sum of true positive rates divided by the sum of false positive rates. The macro average is computed by calculating the AUC for each of the binary cases, and then averaging the results. Accuracy and balanced accuracy metrics were also used to evaluate model performance on the hold-out test set. We compared these results against other available models (Liu et al., 2022, Zhang-James et al., 2023). For feature interpretability, we generated and discussed saliency maps generated by the convolution blocks.

7. Machine Learning Pipeline

The machine learning methodology was modeled after the work of Srinivasan et al., who used both supervised and unsupervised learning techniques to predict suicide attempts amongst women diagnosed with either post-traumatic stress disorder or dissociative identity disorder. Much like the Srinivasan study, ABCD data used in the machine learning pipeline were divided into two primary sections, numerical data and categorical data. Feature and label data were loaded separately before using the SMOTE oversampling technique to account for the heavy imbalance in the control and positive classes. Because the ABCD minority class, positive depression or anxiety diagnoses, are heavily underrepresented in the sample, oversampling was necessary to improve model performance. Using SMOTE, samples were added by creating intermediate points in the existing feature space of the k-nearest neighbors (i.e. k-NN graph) (Srinivasan et al., 2022). This method, although it balances classes, does not provide any additional information for the model. After oversampling, holdout analysis was performed, splitting the dataset into many training and test splits for cross-validation. Recursive feature elimination was used to limit the large number of features present in the ABCD data, resulting in approximately 1000 final features. After preparation, a multiclass random forest model was used to predict diagnosis outcomes.

Machine Learning Pipeline Steps

1. Load ABCD Feature and Label Data
2. Oversampling (SMOTE)
3. Holdout Analysis
4. Recursive Feature Selection
5. Random Forest Model
6. Model Performance Evaluation

8. Machine Learning Evaluation metrics/Model Interpretation

According to findings from Srinivasan et al., we used accuracy and F1 scores to measure the performance of the random forest model on the training and test sets. The top 20 features for predicting disease diagnosis, determined by correlation coefficient value, were visualized using scatter plots and a line of best fit.

Results

1. Study Participants

Created to be a representative sample of the US adolescent population, the ABCD Study is the largest longitudinal study of adolescent brain development imaging dataset at the time of writing. It consists of ~11,800 youth, aged 9–11 years recruited from 21 sites across the US to create a population-level, socio-demographically-diverse sample. Roughly 48% female and 52% male. 52.4% are white. Caucasians make up the largest demographic at 51.9%. Hispanics make up the next largest demographic at 20.5%. African Americans make up 14.7%. Asians make up 2.4%. AIAN, NHPI, and mixed races make up the remainder. We use a filtered subset of the ABCD study of randomly sampled control participants and all available subjects with major depressive disorder diagnosed at the time of the scan. Participant dropout resulted in fewer samples recorded for follow-up visits. For this study, participants were included regardless of study completion status (*ABCD Release Notes 5.1*.).

Event	Population size ()
Baseline	11,868
6-month follow-up	11,389
1-year follow-up	11,220
18-month follow-up	11,083
2-year follow-up	10,973
30-month follow-up	10,228
3-year follow-up	10,336
42-month follow-up	8,449
4-year follow-up	4,754

Table 2: Number of participants included in ABCD study duration.

Characteristic	Category	ABCD n=11,874	ACS 2011-2015	
		%	$\hat{N}$	%
Population	Total	100.0	8,211,605	100.0
Age	9	52.3	4,074,807	49.6
	10	47.7	4,136,798	50.4
Sex	Male	52.2	4,205,925	51.2
	Female	47.8	4,005,860	48.8
Race/Ethnicity	NH White	52.2	4,305,552	52.4
	NH Black	15.1	1,101,297	13.4
	Hispanic	20.4	1,973,827	24.0
	Asian, AIAN, NHPI	3.2	487,673	5.9
	Multiple	9.2	343,256	4.2
Family Income	<$25K	16.1	1,762,415	21.5
	$25K-$49K	15.1	1,784,747	21.7
	$50K-$74K	14.0	1,397,641	17.0
	$75K-$99K	14.1	1,023,127	12.5
	$100K-$199K	29.5	1,685,036	20.5
	$200K +	11.2	558,639	6.8
Family Type	Married Parents	73.4	5,426,131	66.1
	Other Family Type	26.6	2,785,474	33.9
Parent Employment	Married, 2 in LF	50.2	3,353,572	40.8
	Married, 1 in LF	21.9	1,949,288	23.7
	Married, O in LF	1.3	156,807	1.9
	Single, in LF	21.1	2,174,365	26.5
	Single, Not in LF	5.4	577,573	7.0
Region	Northeast	16.9	1,336,183	16.3
	Midwest	20.4	1,775,723	21.6
	South	28.3	3,117,158	38.0
	West	34.4	1,982,541	24.1
Household size	2-3	17.3	1,522,216	18.5
	4	33.5	2,751,942	33.5
	5	24.9	2,085,666	25.4
	6	14.0	1,025,285	12.5
	7+	10.3	826,496	10.1

Table 3: ABCD Baseline Cohort Demographic and Socio-Economic Characteristics (Heeringa & Berglund, 2020)

2. Identification of Depressive Disorders vs. Healthy Controls – Deep Learning

We framed this problem as a binary classification problem. Due to dataset constraints, we ended up with 888 subjects, around three quarters of which belonged to the control group and the latter to the depression group. After applying all previously discussed augmentations, the dataset was split into a training, validation, and testing set following a 0.8/0.1/0.1 split respectively.

The model was fine-tuned using the pre-trained model from (Liu et al., 2022) as an image embedder network. These layers are then frozen, and a classification head was added on top of the frozen layers which is then trained to classify based on the 1024 features returned by the image embedder network. Grid search was used to optimize hyper-parameters (learning rate, batch size, epochs) based on balanced accuracy evaluated on the validation set.

We were unable to find a solution after several days of searching; the model consistently converged to a 'lazy solution' regardless of hyper parameters which is to make a negative prediction regardless of the sample given to the model which gave it a ~75% accuracy (the proportion of control subjects) but is generally considered invalid for clinical use. We then introduced weighted loss for each class such that the weight for each class is inversely proportional to the proportion of that class in the dataset to force a non-lazy solution. IE: loss for a control subject which makes up three-quarters of the dataset would be worth a third of loss for a non-control subject makes up a quarter of the dataset.

The introduction of weighted loss caused the model to instead vary between all negative and all positive predictions, occasionally making mostly negative/positive predictions. We hypothesize that this is due to the classification head being too simple to fit to the feature space being provided by the image-embedder network, at which point we began introducing wider layers as well as increasing the depth of the classification head network.

Initially, the classification head consists of a 1024-unit flattened layer, connected to a 512-unit hidden layer, which then connects to the 2-unit output layer. Notably, there were no activation functions in the classification head aside from SoftMax normalization which is applied at the end. We experimented with several architectures for the network, varying width and depth as well as several activation functions including Tanh, Sigmoid, Rectified Linear Unit (ReLU), and even Leaky ReLU.

These changes were successful at allowing the model to learn a non-trivial solution, but performance remained unviable. The best result was achieved using an eight-layer deep classification head consisting solely of tanh with width varying between 1024 and 512. While the model no longer defaulted to one class over the other and was at least attempting to make predictions based on the provided features, the highest balanced validation accuracy was around ~63% with a true positive rate of ~48% and a true negative rate of ~78%. Given that the model gives a negative prediction, there is a ~79% chance the model is correct, while given that the model gives a positive prediction, there is a ~46% chance the model is correct. Overall, these accuracy results are poor and suggest an inability to fit the data.

3. Identification of Depressive Disorders vs. Healthy Controls – Machine Learning

Due to computational limits, the ABCD sample for the machine learning pipeline was limited to 500 subjects, 441 controls and 59 non-controls. All non-control subjects with existing and complete records for depression diagnoses and structural MRI imaging were included. The remaining subjects were chosen at random.

After applying SMOTE for oversampling, recursive feature elimination revealed 39 features to be optimal for model performance. Because of a higher feature and low sample number, many of the features were redundant.

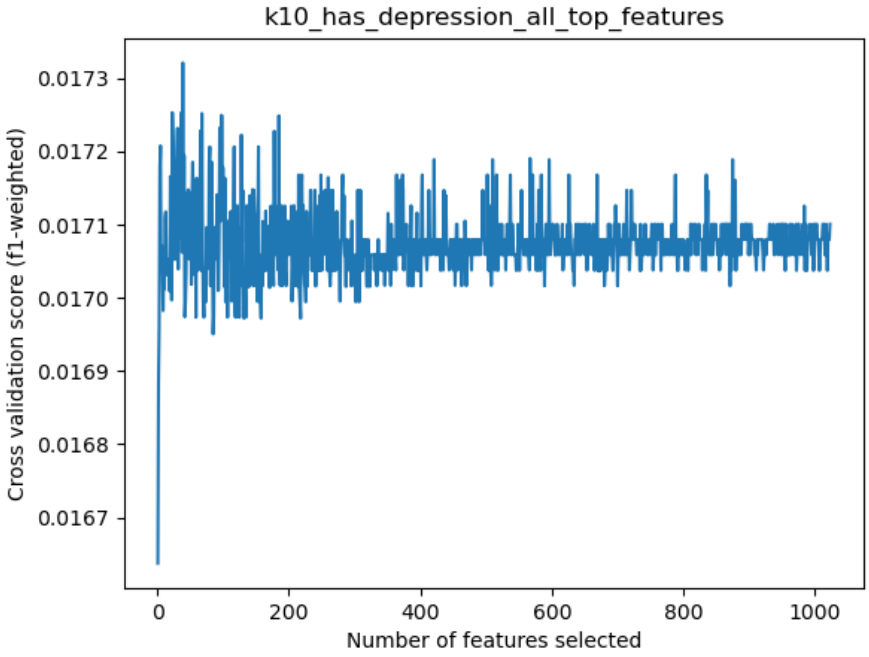

Figure 2: Number of features selected vs. Cross validation score for holdout set, showing optimal features a 39.

Top Random Forest model features by importance were reported and are shown below. Largely, thickness and volume measures were the most indicative of control vs. non-control classification. Cortical thickness of the postcentral gyrus, inferior temporal gyrus, and anterior transverse gyrus were the top three features. Gyri are ridge-like elevations on the cerebral cortex of the brain, surrounded by depressions called sulci. Gyri and sulci create the brain's folded surface, which is believed to increase the brain's surface area and cognitive ability (White et al., 2009). Abnormal gyrification has been cited in literature as a potential biomarker for depression and anxiety in adolescents (Pagliaccio et al, 2020). However, many of the other key brain ROIs cited such as abnormalities in amygdala and putamen volumes were not important features on the model. Additionally, a low F1 score of 0.333 suggests that the model performed poorly against the holdout set, resulting in low precision and low recall.

ABCD Release 5.1 Feature Name	Feature Value
Mrisdp_28	Cortical thickness in mm for left hemisphere cortical Destrieux ROI postcentral gyrus
Mrisdp_111	Cortical thickness in mm for right hemisphere cortical Destrieux ROI inferior temporal gyrus

Mrisdp_33	Cortical thickness in mm for left hemisphere cortical Destrieux ROI anterior transverse temporal gyrus
Smri_thick_cdk_insularh	Cortical thickness in mm of APARC ROI rh-insula
Mrisdp_64	Cortical thickness in mm for left hemisphere cortical Destrieux ROI orbital sulci
Smri_thick_cdk_pericclh	Cortical thickness in mm of APARC ROI lh-pericalcarine
Smri_thick_cdk_periccrh	Cortical thickness in mm of APARC ROI rh-pericalcarine
Mrisdp_58	Cortical thickness in mm for left hemisphere cortical Destrieux ROI superior occipital sulcus and transverse occipital sulcus
Smri_thick_cdk_postcnlh	Cortical thickness in mm of APARC ROI lh-postcentral
Smri_vol_scs_pallidumlh	Volume in mm^3 of ASEG ROI right-pallidum

Table 4: Top 10 most important features of Random Forest model and feature explanation according to the ABCD Release 5.1 Data Dictionary.

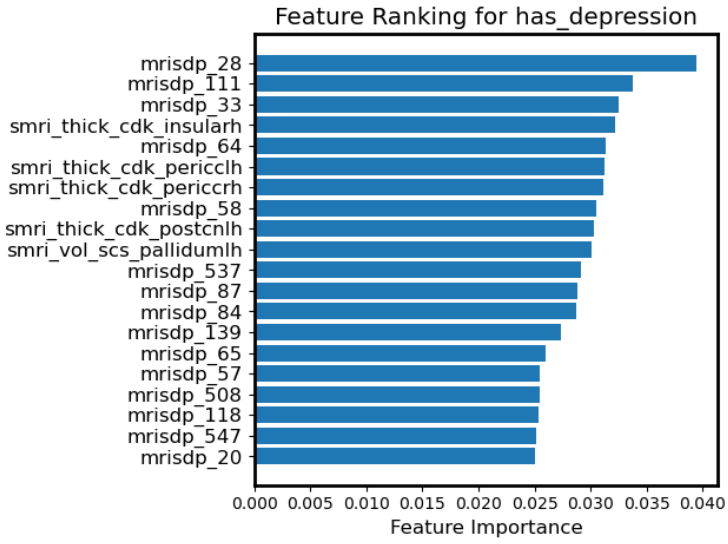

Figure 3: Top 20 Random Forest features by feature importance.

Discussion

Several explanations exist for poor model performance results—we believe the most likely explanation is that the neuroanatomical differences between presented by subjects with Alzheimer's Disease from the Liu et al. model compared with subjects with depression from ABCD are too distinct from each other, hence the features learned by the image-embedder network are not applicable in MDD. From a neuroanatomical perspective, AD is a neurodegenerative disease that leads to very apparent structural changes. Meanwhile, MDD is much more subtle and varied by comparison. This is not to say that MDD does not cause neuroanatomical changes. Numerous studies support the idea that long-term depression impacts development of various brain regions. Whittle et al. for example reported attenuated growth in the hippocampus and putamen associated with onset of depression in adolescents. A more general review of MRI studies in MDD by Zhuo, Chuanjun, et al found many sources reported consistent structural changes due to MDD. This is corroborated by our own analysis of the structural scans as well. Welch's t-test for difference of mean volumes was ran for 818 regions with significance level p=0.05. Out of those regions we were able to reject the null hypothesis (equal means across groups) for 253 regions out of the 818 regions tested

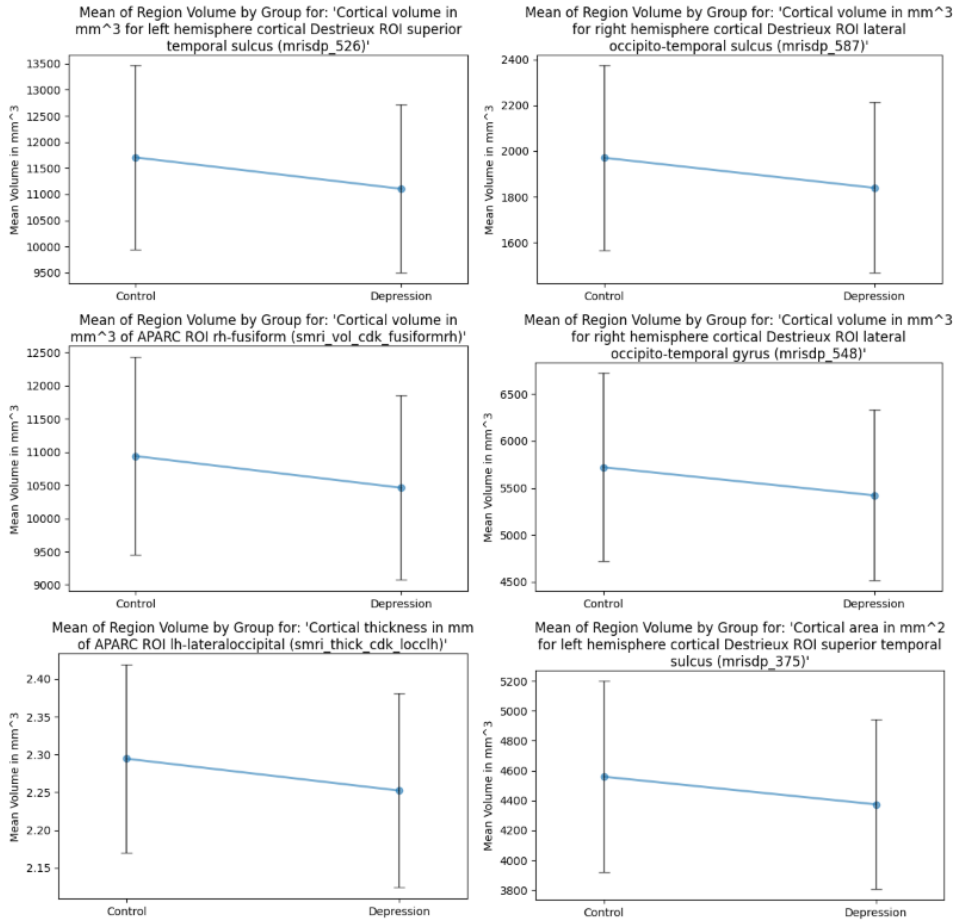

Figure 4: *Six examples of statistically significant volume differences, in all statistically significant regions, we found lower region volume for the MDD group*

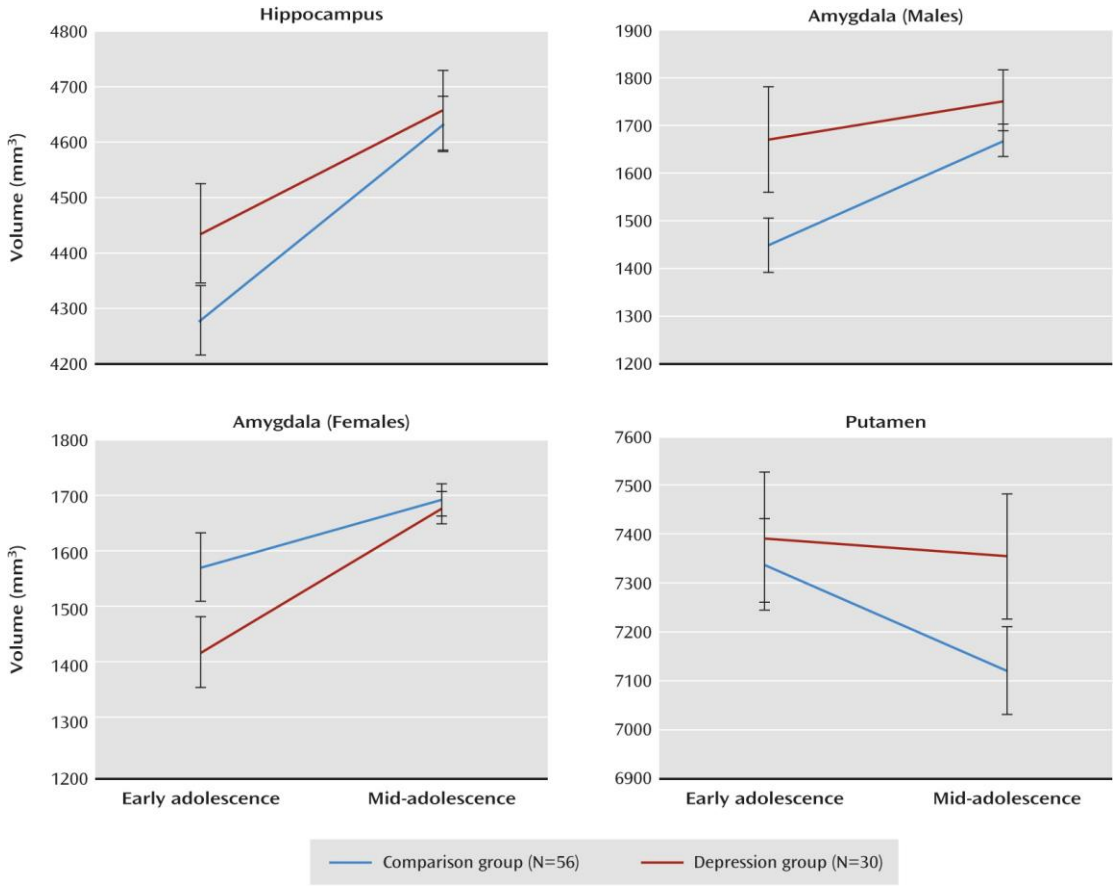

Figure 5: *From "*Structural Brain Development and Depression Onset during Adolescence: A Prospective Longitudinal Study." *by Whittle, Sarah, et al.*

However, a statistically significant structural difference in certain regions does not mean those regions make robust predictors by themselves. In both our analysis and findings from Whittle et al., the mean volumes are statistically distinct between the MDD and control group in certain regions—yes, but there is also a significant overlap in the distribution of volumes in those regions. As a very blunt example: while control subjects demonstrate higher average volume in the left hemisphere, superior temporal sulcus than MDD subjects (~11706 mm^3 vs. ~11105 mm^3 respectively), roughly over 35.46% of the sampled MDD subjects has higher volume in this region than the control group's mean and 36.62% of control subjects has a lower volume in this region than the MDD group's mean. Another thing to note from the Whittle, et al. visualization is that these distributions change with age, which we were unable to adjust for, further reducing their viability as discriminators. It could be the cased that these regions viewed in conjunction contextually with age would lead to a more robust discriminator, but incorporating age information into the embeddings is an area of future work.

The structural changes manifested by MDD are significantly more subtle than those manifested by AD and in different regions. Additionally, there is an age difference in patients with AD and our sample of adolescent patients with MDD. These differences combined make it highly likely that crucial cortical information that would have been useful for discriminating depression from control subjects is lost after passing the neuroimages through the pre-trained AD

neuroimage embedder. It is also possible that the method of image preprocessing removed critical cortical data for the model. The Clinica software removes part of the brain cortex during preprocessing that could potentially improve model performance. Cortical abnormalities have been found in patients with depression and anxiety (Strawn et al., 2015).

From a deep learning perspective, this is supported by the fact that during our training, the balanced training accuracy remained consistently low (less than <70% throughout), indicating a severe inability to fit data to the labels. Due to computational power constraints, we were unable to experiment with partially unfreezing the image embedder network to experiment with training low-level CNN feature representations as well. Even without the computational constraint, however, the size of the available dataset would be too small regardless to properly train a CNN even partially. However, this remains an unexplored area that future works could tackle. Later feature maps in a CNN are generally more specialized towards the problem domain while earlier feature maps tend to be more generic. This means that while the features learned by the pre-trained network at the final layer offer an effective representation for the ADNI dataset, it might not offer a useful representation for our depression neuroimaging classification problem where subjects are significantly younger, and the nature of the disorder is completely different. Therefore, retaining earlier feature maps and training the later stages on our specific problem with a larger dataset may still yield compelling results, but alas this is to be explored.

Another observation made was that ReLU and sigmoid generally performed worse than Tanh. The common denominator between these two activation functions is that they are non-zero centered activation functions. These functions require incoming data to be zero-centered, otherwise training becomes very inefficient as the direction of gradient descent becomes restricted to either the direction of increasing all parameters or of decreasing all parameters rather than in any arbitrary direction in the parameter space. The figure below illustrates an example of a gradient surface where descent is restricted to either the direction which increases all parameter values or decreases all parameter values—a sort of upside-down saddle in which different models with different weight initializations converge to different local minimas in the valley and once they struggle to converge to any better solution.

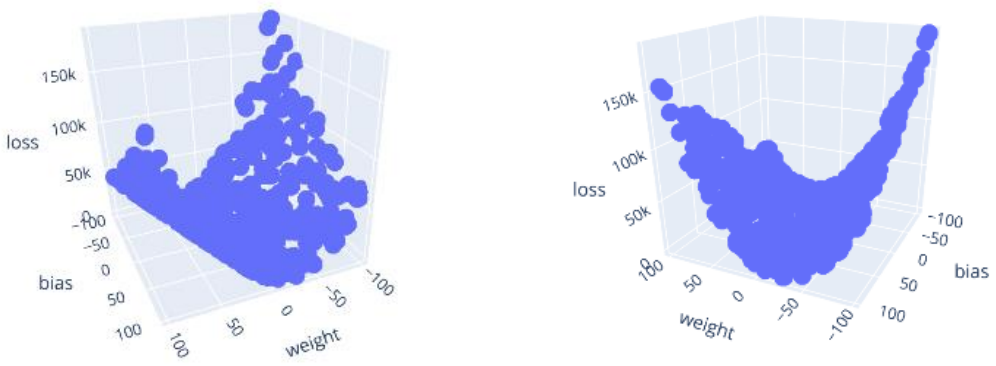

Figure 6,7: *A simplified visualization of what a poor gradient surface caused by non-zero centered data might look like. Gradient descent causes models to rapidly descend into the 'valley' and once reached, descent becomes extremely inefficient to the point of being unviable, causing the model to converge at unsatisfactory local minima solutions.*

It may be the case that the distribution of the dataset mapped to the feature space by the image embedder network is non-zero centered, causing highly inefficient gradient descent. In the case of ReLU we also observed the 'dying ReLU' problem in several instances in which the model gets stuck with a trivial learned representation and is unable to get out of it because the gradient is zero. We assumed this was an issue with the learning rate being too high and lower learning rates did ameliorate the issue to a degree as did using Leaky ReLU instead.

Future Directions

Due to poor model performance, there are many possibilities to improve upon this research. Without computational constraints, unfreezing the image embedder network to experiment with training low-level CNN feature representations would be useful in determining whether the Liu et al model is translatable from Alzheimer's to depression and anxiety. A proper replication of the model using Alzheimer's data would be necessary. Additionally, experimenting with non-preprocessed images or a different preprocessing pipeline may determine if poor performance was due to data quality issues. It may also be useful to attempt separate analyses using functional MRI data as functional brain connectivity has been noted to be affected in adolescents with mental health disorders in literature. For the complementary machine learning work, separate models for thickness, surface area, and volume metrics could be created to restrict regions of interest. In this study, only a Random Forest model was implemented. Other classification models including Logisitic Regression and Support Vector Machine (SVM) have been proven effective and should be applied to this research.

Data Availability

All imaging data obtained from the ABCD study is available at no cost following approval of a Data Access Request (DAR) from the National Institute of Mental Health (NIMH). To obtain access to ABCD data, users need to create a National Institute of Mental Health Data Archive (NDA) account. From this account, users may complete the access request instructions on the Data Permissions dashboard which includes requesting electronic access through an automated Data Use Certification (DUC) or by submitting a signed copy to NDAHelp@mail.nih.gov. To obtain ABCD access, all users specific by the DUC must have a research-related need to access the data and be associated with an NIH-recognized research institution that has active Federal wide Assurance (FWA). Once approved by the ABCD Data Access Committee, users have full ABCD access for one year through the NDA.

For our study, all data were taken from ABCD Release 5.1. Easy access to all tabulated variable locations and descriptions can be located through the ABCD Data Dictionary (https://data-dict.abcdstudy.org/?). Imaging data must be downloaded using the NDA Download Manager after adding desired data to a customized Data Package on the user's profile.

The deep learning pipeline was performed on deidentified data, so IRB Review was not required. The 3D Convolutional Neural Network pipeline from which this study was modeled can be found at https://github.com/NYUMedML/CNN_design_for_AD.

References

ABCD Data Dictionary (Release 5.1). (n.d.). Retrieved March 18, 2024, https://data-dict.abcdstudy.org/?

ABCD Release Notes 5.1. (2024, March 7). https://wiki.abcdstudy.org/release-notes/start-page.html

ABCD Study. (n.d.). ABCD Study. Retrieved March 18, 2024, https://abcdstudy.org/

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., & Lladó, X. (2019). Deep convolutional neural networks for brain image analysis on Magnetic Resonance Imaging: A Review. *Artificial Intelligence in Medicine*, *95*, 64–81. https://doi.org/10.1016/j.artmed.2018.08.008

CDC. (2022, June 3). *Data and Statistics on Children's Mental Health | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/childrensmentalhealth/data.html

DSM. (2013). Retrieved September 25, 2023, from https://www.psychiatry.org:443/psychiatrists/practice/dsm

FreeSurferWiki—Free Surfer Wiki. (n.d.). Retrieved March 18, 2024, from https://surfer.nmr.mgh.harvard.edu/fswiki

Gai, Q., Chu, T., Che, K., Li, Y., Dong, F., Zhang, H., Li, Q., Ma, H., Shi, Y., Zhao, F., Liu, J., Mao, N., & Xie, H. (2022). Classification of major depressive disorder based on integrated temporal and spatial functional mri variability features of Dynamic Brain Network. *Journal of Magnetic Resonance Imaging*, *58*(3), 827–837. https://doi.org/10.1002/jmri.28578

Gale, C. K., & Millichamp, J. (2016). Generalised anxiety disorder in children and adolescents. *BMJ Clinical Evidence*, *2016*, 1002.

Garber, J., & Weersing, V. R. (2010). Comorbidity of Anxiety and Depression in Youth: Implications for Treatment and Prevention. *Clinical Psychology: A Publication of the Division of Clinical Psychology of the American Psychological Association*, *17*(4), 293–306. https://doi.org/10.1111/j.1468-2850.2010.01221.x

Heeringa, S. G., & Berglund, P. A. (2020). A Guide for Population-based Analysis of the Adolescent Brain Cognitive Development (ABCD) Study Baseline Data [Preprint]. Neuroscience. https://doi.org/10.1101/2020.02.10.942011

Li, Q., Dong, F., Gai, Q., Che, K., Ma, H., Zhao, F., Chu, T., Mao, N., & Wang, P. (2023). Diagnosis of major depressive disorder using machine learning based on multisequence mri neuroimaging features. *Journal of Magnetic Resonance Imaging*. https://doi.org/10.1002/jmri.28650

Liu, S., Masurkar, A. V., Rusinek, H., Chen, J., Zhang, B., Zhu, W., Fernandez-Granda, C., & Razavian, N. (2022). Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-20674-x

Pagliaccio, D, Alqueza, K, Marsh, R, Auerbach, R. (2020). Brain Volume Abnormalities in Youth at High Risk for Depression: Adolescent Brain and Cognitive Development Sudy. *Journal of the American Academy of Child & Adolescent Psychiatry. 59(10)* https://doi.org/10.1016/j.jaac.2019.09.032

Pettitt, R. M., Brown, E. A., Delashmitt, J. C., & Pizzo, M. N. (n.d.). The Management of Anxiety and Depression in Pediatrics. *Cureus*, *14*(10), e30231. https://doi.org/10.7759/cureus.30231

Protocols - ABCD Study. Adolescent Brain Cognitive Development. (2024). https://abcdstudy.org/scientists/protocols/

Saragosa-Harris, N. M., Chaku, N., MacSweeney, N., Guazzelli Williamson, V., Scheuplein, M., Feola, B., Cardenas-Iniguez, C., Demir-Lira, E., McNeilly, E. A., Huffman, L. G., Whitmore, L., Michalska, K. J., Damme, K. S., Rakesh, D., & Mills, K. L. (2022). A practical guide for researchers and reviewers using the ABCD Study and other large longitudinal datasets. Developmental Cognitive Neuroscience, 55, 101115. https://doi.org/10.1016/j.dcn.2022.101115

Shorey, S., Ng, E. D., & Wong, C. H. J. (2022). Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis. *British Journal of Clinical Psychology*, *61*(2), 287–305. https://doi.org/10.1111/bjc.12333

Srinivasan, S., Harnett, N. G., Zhang, L., Dahlgren, M. K., Jang, J., Lu, S., Nephew, B. C., Palermo, C. A., Pan, X., Eltabakh, M. Y., Frederick, B. B., Gruber, S. A., Kaufman, M. L., King, J., Ressler, K. J., Winternitz, S., Korkin, D., & Lebois, L. A. M. (2022). Unravelling psychiatric heterogeneity and predicting suicide attempts in women with trauma-related dissociation using artificial intelligence. *European Journal of Psychotraumatology*, *13*(2), 2143693. https://doi.org/10.1080/20008066.2022.2143693

Walter, H. J., Abright, A. R., Bukstein, O. G., Diamond, J., Keable, H., Ripperger-Suhler, J., & Rockhill, C. (2023). Clinical Practice Guideline for the Assessment and Treatment of Children and Adolescents with Major and Persistent Depressive Disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, *62*(5), 479–502. https://doi.org/10.1016/j.jaac.2022.10.001

White, T, Su, S, Schmidt, M, Kao, C, Sapiro, G. The Development of Gyrification in Childhood and Adolescence. *Brain Cogn. 71(1). https://doi.org/10.1016/j.bandc.2009.10.009*

Whittle, Sarah, et al. “Structural Brain Development and Depression Onset during Adolescence: A Prospective Longitudinal Study.” *American Journal of Psychiatry*, 1 May 2014, ajp.psychiatryonline.org/doi/10.1176/appi.ajp.2013.13070920.

Xiang, Q., Chen, K., Peng, L., Luo, J., Jiang, J., Chen, Y., Lan, L., Song, H., & Zhou, X. (2022). Prediction of the trajectories of depressive symptoms among children in the adolescent brain cognitive development (ABCD) study using machine learning approach. *Journal of Affective Disorders*, *310*, 162–171. https://doi.org/10.1016/j.jad.2022.05.020

Zhang-James, Y., Razavi, A. S., Hoogman, M., Franke, B., & Faraone, S. V. (2023). Machine learning and MRI-based diagnostic models for ADHD: Are we there yet? *Journal of Attention Disorders*, *27*(4), 335–353. https://doi.org/10.1177/10870547221146256

Zhuo, Chuanjun, et al. “The Rise and Fall of MRI Studies in Major Depressive Disorder.” *Nature News*, Nature Publishing Group, 9 Dec. 2019, www.nature.com/articles/s41398-019-0680-6#Sec2.

Zubrick, S. R., Hafekost, J., Johnson, S. E., Sawyer, M. G., Patton, G., & Lawrence, D. (2017). The continuity and duration of depression and its relationship to non-suicidal self-harm and suicidal ideation and behavior in adolescents 12–17. *Journal of Affective Disorders*, *220*, 49–56. https://doi.org/10.1016/j.jad.2017.05.050