Abstract

Title: ANALYSIS OF WHOLE-BRAIN RESTING-STATE MRI USING MULTI-LABEL DEFORMABLE OFFSET NETWORKS AND SEGMENTATION BASED ATTENTION WITH EXPLORATIONS INTO THE ETHICAL IMPLICATIONS OF ARTIFICIAL INTELLIGENCE IN CLINICAL PSYCHIATRY SETTINGS AND CARE

Vatsal Agarwal, Evan Ayoroa, Ryerson Burdick, Aravind Ganeshan, Madhava Paliyam, Sam Wood, Caitlin Lee, Sepehr Akhtarkhavari, Shika Inala, Sagar Matharu, Neelesh Mupparapu

Directed by: Dr Anil Deane
Institute for Physical Science and Technology, University of Maryland

Due to the poor understanding of the underlying biological mechanisms of psychiatric disorders, diagnoses rely upon symptomatic criteria and clinicians' discretion. Reviews of these criteria have revealed issues of heterogeneity, over and under specificity, and symptom overlap between disorders. Deep learning provides a method to produce quantifiable diagnostic labels based upon biological markers such as specific features of brain anatomy or functionality. In practice, these methods fail to indicate how a particular result was determined, raising major obstacles for clinical implementation.To improve the efficiency and interpretability of existing deep networks, we have developed a novel atlas-based attention module to more easily capture global information across different areas of brain function. Our model can be extended to symptom level classification using NIMH data to give clinicians usable information outside of broad disorder classification. We have compared our model against leading 3D deep learning frameworks and have shown that our novel atlas-based attention module achieves 88% F1 and 91% accuracy on the UCLA Consortium for Neuropsychiatric Phenomics dataset. We have embedded our model with elements like deformable convolutions, gradient activation visualizations, and occlusion testing to show model attention and function. In addition to the lack of explainability, addressing the ethical issues surrounding clinical implementation of artificial intelligence is necessary before usage can become a reality. We identified a series of regulatory recommendations to address pertinent ethical concerns of equity and bias during both model development and clinical usage. We propose a standardized protocol for developing a clinical reference standard, the development of diversity reports regarding data used by models, and regulation of usage scenarios to reduce contextual bias.

Analysis of Whole-Brain Resting-State MRI Using Multi-label Deformable Offset Networks and Segmentation Based Attention with Explorations into the Ethical Implications of Artificial Intelligence in Clinical Psychiatry Settings and Care

By

Team MIND

Vatsal Agarwal, Evan Ayoroa, Ryerson Burdick, Aravind Ganeshan, Madhava Paliyam, Sam Wood, Caitlin Lee, Sepehr Akhtarkhavari, Shika Inala, Sagar Matharu, Neelesh Mupparapu

Thesis submitted in partial fulfillment of the requirements of the
Gemstone Honors Program, University of Maryland 2022

Mentor: Dr. Anil Deane

Advisory Committee:
Dr. Michele Ferrante
Dr. David Benedek
Dr. Abhinav Shrivastava
Mr. Bradley Cunningham

Acknowledgements

We would like to thank our mentor Dr. Anil Deane, for all the insightful advice and mentorship he has given our team in the past 3 years. We would also like to thank our librarian Jordan Sly for his invaluable assistance. We would like to thank our discussants, Dr. Michele Ferrante, Dr. David Benedek, Dr. Abhinav Shrivastava, and Mr. Bradley Cunningham for their expert critique and guidance throughout our research. Additionally, we would like to thank the gemstone staff Dr. Coale, Dr. Lovell, Dr. Skendall, Dr. Hill, and Dr.Tobin for their support.

Contents

1 Introduction

1.1 Shortcomings of Psychiatric Diagnosis

Mental illnesses are among the most prevalent health conditions worldwide. In 2018, it was estimated that nearly 20% of adults in the U.S. (47.6 million people) experienced some form of mental illness (*Mental Health By the Numbers*, n.d.). Serious mental illnesses affect approximately 5% of U.S. adults (11.4 million people) (*Mental Health By the Numbers*, n.d.). For the majority of the past century, psychiatric diagnoses have relied exclusively upon categorical symptomatic criteria established in guides such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and International Classification of Diseases (ICD-10) (Schultze-Lutter et al., 2018; Suris et al., 2016). The DSM provides clinicians with an explicit set of symptom requirements for each disorder. These criteria are established in a polythetic structure, with disorder categories consisting of multiple symptom clusters that must be fulfilled by meeting some minimum number of symptoms within each cluster (Young et al., 2014). This has the purpose of providing flexibility to the diagnosis in order to encompass a variety of patient presentations. While having these criteria does improve clinician reliability, heterogeneity within and across disorders exemplify inconsistencies throughout criteria.

Heterogeneity across disorders manifests as significant symptom overlap across disorder criteria. For example, hallucinations are part of criteria for psychotic disorders such as schizophrenia, but also appear in the criteria for bipolar and related disorders, and PTSD (Allsopp et al., 2019, p. 19). This leads to the relatively high frequency of psychiatric comorbidities, i.e. the presence of two or more disorders in one patient (Feczko et al., 2019; van Loo & Romeijn, 2015). Van Loo and Romeijn estimate that up to 45% of patients per year meet the minimum requirements for multiple comorbid disorders (van Loo & Romeijn, 2015). Symptom overlap and comorbidity suggests that the divisions between disorders are not as clear as the disorder categorizations suggest. Without quantitative tests or more complete understanding of disorder physiology however, it is not evident if this is an artifact of the DSM's classification scheme, or if there is a deeper biological commonality between frequently comorbid disorders (van Loo & Romeijn, 2015).

Heterogeneity within disorders means that a single disorder diagnosis can come from multiple different symptom combinations. In over half of DSM-5 disorders, two patients can be diagnosed with the same disorder while sharing no common symptoms (Allsopp et al., 2019, p. 1). For some disorders, the degree of heterogeneity is especially excessive. For example, Galatzer-Levy and Bryant calculated that 227 unique combinations of symptoms exist for a major depressive episode, 23,442 combinations exist for panic disorder, and 636,120 combinations exist for post-traumatic stress disorder (PTSD). Furthermore, if comorbid disorders are taken into account, the number of possible presentations of PTSD and its comorbidities jumps to over one quintillion (Young et al., 2014). The existence of internal heterogeneity

suggest that these disorders consist of subtypes or exist on continuums, with differences in causal mechanisms underlying these variations (Allsopp et al., 2019, p. 21; Feczko et al., 2019)

Another shortcoming of symptomatic criteria is that diagnoses significantly depend upon clinician judgment. A consequence of this is that mental health diagnosis is especially at risk for being negatively affected by unconscious thoughts and feelings, known as implicit bias (Merino et al., 2018, p. 723). Implicit biases are especially likely to impact patients from minority groups. Snowden identifies two main types of bias: overpathologizing bias, in which unfamiliar behaviors of minorities are misidentified as disorder symptoms, and minimization bias, in which actual disorder symptoms are ignored (Snowden, 2003). These biases likely contribute to the overdiagnosis and underdiagnosis of certain disorders, respectively. Minority groups are more likely to be underdiagnosed with affective disorders such as depression and bipolar disorder, and overdiagnosed with psychotic disorders such as schizophrenia and schizoaffective disorder (Merino et al., 2018, p. 724; Schwartz & Blankenship, 2014, p. 134). In particular, Barnes found that African Americans in state hospitals are almost five times more likely than white Americans to be diagnosed with schizophrenia. Likewise, Chien et al. state that the disproportionate diagnosis of African Americans holds despite the epidemiological incidence of schizophrenia being equal across racial groups, suggesting that there is some factor within clinical diagnosis contributing to the disparity.

Improvements to psychiatric diagnosis should be made to address the shortcomings of current disorder criteria. However, the lack of understanding of brain functioning means that the underlying biological causes of these disorders are also unknown, thereby forcing the continued dependence on symptomatic descriptions of disorders. Working towards a more descriptive definition of psychiatric disorders (ie. subtypes, pathophysiology, stage/severity, biological predisposition) through biomarkers will help physicians move past the currently flawed standards (Waszkiewicz, 2020). In addition, this will allow for the development of treatments targeted at specific dysfunctional systems (Kendell & Jablensky, 2003, p. 4). Progress in developing and establishing this understanding has been slow due to the complexities of the brain. Disorders are likely caused by a combination of biological factors, e.g. neuroanatomy, neurochemistry, and genetics (Kendell & Jablensky, 2003). Furthermore, a review of current techniques reveals the key fact that confounding variables such as neurological disorders, environmental factors, psychoactive substance use, comorbidities, and medication reduce the efficacy of measurable markers (Waszkiewicz, 2020).

1.2 Neuroimaging for Biomarker Identification

Our search for descriptive biomarkers began with the assumption that evidence for psychiatric disorders exists in the brain. While brain imaging is more commonly used to diagnose trauma and measure brain function, trends in populations suffering from psychiatric disorders such as altered brain volume and decreased brain activity raises the question of identifiability via neuroimaging (Mayo Clinic, 2021; Waszkiewicz, 2020). It is understood that morphological changes occur in regions tied to emotional and cognitive function. For example,

depression leads to decreased volume in the hippocampus, prefrontal cortex, orbitofrontal cortex, and anterior cingulate cortex (Wilczyńska et al., 2018). A meta-analysis quantified this change, finding an average reduction in hippocampal volume of 8 and 10 percent for the left and right sides, respectively, in patients with a history of depression (Videbech and Ravnkilde, 2004). In addition, postmortem microscopic analysis found abnormal variations in cell count and density in those who suffered from mood disorders (Wilczyńska et al., 2018). With measurable differences found in affected subjects, we move forward with the assumption that features can be extracted from imaging data to help classify psychiatric disorders.

Magnetic resonance imaging (MRI) is a noninvasive method for doctors to produce high resolution three-dimensional images of the brain. It has been used for almost 40 years in the diagnosis of psychiatric disorders to help classify potential structural differences and rule out underlying brain diseases (Falkai et al., 2018). The prevalence of neuroimaging for better identification of psychiatric disorders gives us the opportunity to analyze images based on their salient features. In the world of computer vision, this is broadly defined as clustering and classification. The most applicable traditional computer vision methods are detector and descriptor schemes, such as a Scale-invariant feature transform (SIFT) (Georgiou et al., 2020). SIFT relies on finding and describing keypoints, corners and strong edges, to help quantitatively define a class of images. This method has been tested on ultrasounds, similar to an MRI, to automate registration and stitch together 3D panoramas (Ni et al., 2009). However, our problem requires identification of minute differences in the volume, density, or even structure of the brain. A scale invariant technique is able to match features across different images using the corresponding descriptors, but it is up to the user to determine which features best describe a class (O'Mahony, 2019).

1.3 Deep Learning

As opposed to traditional techniques, deep learning (DL) accurately finds the salient features of a class without user input or trial and error (O'Mahony, 2019). While we will cover more in depth descriptions of specific DL algorithms, DL can broadly be defined as "a multilayered representation of input data" (Georgiou et al., 2020). A representation of input data in a higher dimensional space allows analysis to be done and results to be obtained that stretch beyond detector and descriptor schemes and lend well to more descriptive quantitative definitions of psychiatric disorders.

There are numerous examples of studies using neural networks and brain imaging to better classify psychiatric disorders in specific populations. Youth ADHD was studied using over four thousand MRI scans and a multi layer perceptron (Zhang-James et al., 2020). This basic network was run on 151 variables produced after segmenting the images during preprocessing. The results pointed to statistically significant subcortical volumetric reduction, cortical thinning, and reduced brain surface area in children with ADHD. This work led to insights beyond classification and diagnosis, namely proof for the reduction in severity of ADHD with aging. Similarly, with a dataset of only 174 functional MRI scans of subjects diagnosed with

schizophrenia, researchers found that they could achieve a classification accuracy of over 80% by working with only a small number of features (0.5% of all features) generated by their network (Kalmandy et al., 2019). Information outside of pure classification can be extracted from imaging data when using machine learning techniques. These studies show that well defined input data (ie. through segmentation) combined with network defined features not only produces accurate results, but gives researchers the ability to analyze information encoded in specific brain regions.

Most DL based AI systems are analogous to a black box; i.e., the method with which they reach their results are largely, if not completely, unknown. Increased fidelity in explanation and model attention will allow for more detailed analysis of outputs and translate into greater user understanding (Alqaraawi et al. 2020). Explainability often translates to a visual interpretation of the model output, giving the user an idea of the features a model pinpoints and allows users to take a look under the hood. Various methods of explanation can quantitatively verify model robustness. Furthermore, the addition of a DL model as a supplementary tool for clinicians could improve patient care (Tschandl et al., 2020). However, if a DL model were to be implemented in the clinical setting, explainability of results will be crucial for establishing trust with both the patient and clinician.

We propose a deep learning model structure implementing both supervised and unsupervised networks to identify anatomical and functional correlates of various psychiatric disorders and their symptoms. The addition of post-hoc explainability through numerous visualization techniques will help facilitate a better understanding of model behavior so its attention can be used by researchers to better understand the brain, as well as provide a proof of concept of the potential utility of a deep learning model as a supplementary tool for diagnosis. We also propose methods of addressing concerns of bias and inequity from the deployment of artificial intelligence systems in general within clinical settings. The computational foundations of our work and two iterations of our model design are explored in Chapter 2. Our interpretability integrations and visualizations can be viewed in Chapter 3. Finally, the ethical framework and recommendations can be found in Chapter 4.

2 Computational Model Design

2.1 Computational Foundations

2.1.1 Deep Neural Networks

Deep neural networks (DNNs) are a class of machine learning models derived from the perceptron algorithm (Rosenblatt, 1958). Unlike the perceptron and its relative, the multilayer perceptron, deep neural networks are characterized by many layers and a large number of parameters. Each layer in the network takes a vector as an input and computes an output vector

as a nonlinear function of the inputs and a set of tunable parameters called weights (LeCun et al., 2015). These layers are stacked sequentially, with the output of one layer acting as the input to the next, allowing the network to perform increasingly complex manipulations of the input features as its depth increases. The weights of the network are adjusted through a process called backpropagation. Generally, in the case of supervised learning, backpropagation involves adjusting the weights iteratively to minimize the difference between the network's actual output and the expected output. In recent years, deep neural networks have been applied successfully in many areas of machine learning, for example, computer vision, natural language processing, anomaly detection, and structured output learning.

2.1.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a particular type of DNN that are well-suited for image processing. Images normally present a challenge for DNNs due to their relative high-dimensionality, meaning that vanilla fully-connected DNNs often require an impractically large number of weights per layer in order to be able to do any useful image processing. CNNs avoid this challenge through a weight sharing scheme, exploiting the fact that many significant image features are translation-invariant. This approach is loosely inspired by the discovery of local receptive fields in the human retina. Instead of having a corresponding weight for every input-output pair, CNNs implement a set of filters, or kernels, that store weights representing a small portion of the input image. These filters are then "convolved" across the entirety of the input image, such that the same weights are applied in multiple locations.

This weight-sharing scheme significantly reduces the total number of parameters required for image processing, thereby allowing for deeper, more complex networks. Consequently, CNNs have emerged as a leading approach for many computer vision tasks, beginning with the landmark AlexNet paper (Krizhevsky et al., 2017) which achieved what was, at the time, state-of-the-art performance on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). Since then, many models achieving state-of-the-art performance on the ILSVRC benchmark (Russakovsky et al., 2015) have been variants of CNNs.

2.1.3 The ResNet Architecture

Inspired by the observation that deep CNNs tend to perform better as their depth increases (Szegedy, Wei Liu, et al., 2015), ResNets use residual learning to improve the training efficiency of very deep network architectures (He et al., 2015). Unlike vanilla CNN modules, the output of residual layers is taken as the sum of a vanilla CNN layer output and the unaltered input signal. This simple modification mitigates the negative impacts of two well-known problems with the training of deep neural networks. The first is the vanishing gradient problem, which refers to the fact that in traditional backpropagation, error signals from early layers decrease exponentially with network depth. In the worst case, this can result in weight updates for early layers in very deep networks being negligibly small, drastically reducing training

efficiency (Bengio et al., 1994). The second problem concerns the fact that increasing network depth can often lead to saturation and ultimately a degradation of performance (He et al., 2015).

First proposing the ResNet architecture in 2015, He et al. demonstrated that the use of residual layers as opposed to vanilla CNN layers allowed for the effective training of much deeper networks than previously tested (He et al., 2015). Identity operations, as seen in Figure 1, in residual layers meant that there were less vanishing gradients in the deeper layers of the network. Their deep ResNet-based architectures achieved first place on the 2015 ILSVRC classification, ImageNet detection, COCO object detection, and COCO object segmentation tasks. Given their utility and simplicity of implementation, residual connections such as those used in ResNet have become a critical component in many recent deep image processing network architectures.

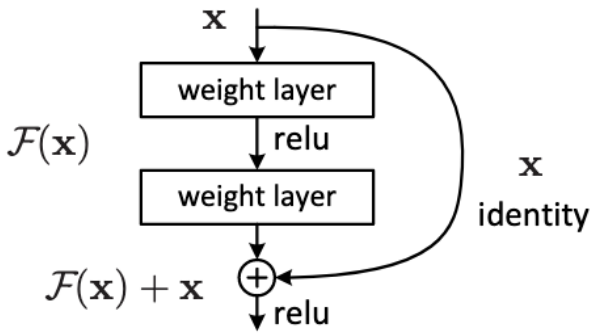

Figure 1. Mathematical operations describing a residual layer in deep networks. The identity operator in addition to the normal weight layer helps with reducing vanishing gradients. Image from He et al.

2.1.4 The U-Net Architecture

U-Net is a convolutional network architecture proposed specifically for biomedical image segmentation (Ronneberger et al., 2015). Given the relatively limited availability of labeled biomedical imaging data, U-Net was designed to leverage data augmentation techniques to train from small imaging datasets (tens of images as opposed to thousands). As seen in Figure 2, the network is organized into two sequentially-connected pathways. First, the contracting pathway downsamples the input image using traditional convolutional and max pooling layers in order to capture image context. Once the contracting pathway performs this feature extraction, the symmetric expanding pathway upsamples the latent image representation using up-convolution layers until the output matches the original input in the height and width dimensions. A final convolution layer is used to assign labels on a pixel-by-pixel basis, effectively producing a

segmentation map of the original image. The original U-Net architecture achieved best performance on the ISBI challenge for the segmentation of neuronal structures in 2015.

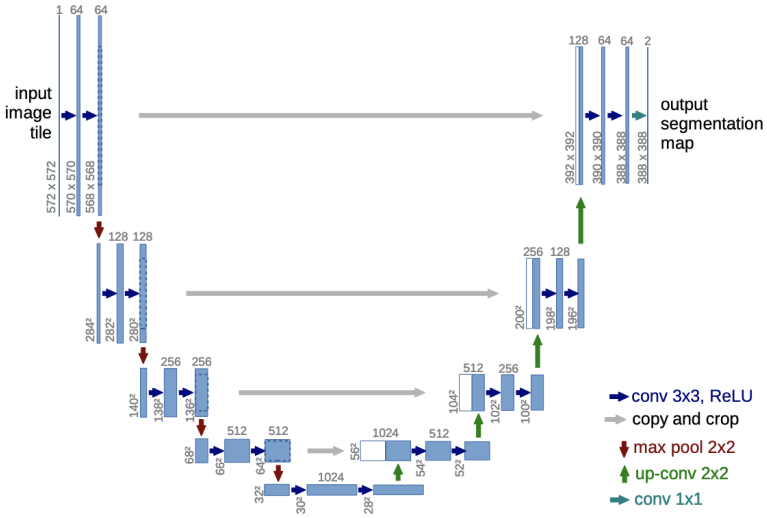

Figure 2. Dimensions and operations done using U-Net architecture for image segmentation. Image from Ronneberger et al.

2.1.5 Attention Mechanisms

Attention is a powerful and widely-used machine learning technique. Generally, attention mechanisms work by enhancing certain parts of the input signal and diminishing other parts, effectively increasing the impact of relevant portions of the input on the output and decreasing that of irrelevant parts. Unlike standard weights, which are fixed at runtime, attention parameters are computed at runtime as a function of the input data. This feature has led attention parameters to sometimes be referred to as "soft weights."

Although attention-like mechanisms have existed in machine learning since at least the 1990's, they gained significant popularity in 2017 with the creation of the Transformer network for natural language processing (Vaswani et al., 2017). The Transformer replaced recurrent components entirely with attention mechanisms for sequence learning, improving performance over previous encoder-decoder approaches while significantly increasing model parallelizability and decreasing overall training time. There is however a downside with the quadratic memory cost. Since then, attention and attention-like mechanisms have been effectively used for a variety of machine learning tasks, including neural machine translation (Tang et al., 2018), image classification (Guo et al., 2019; Ramachandran et al., 2019; Srinivas et al., 2021; Zhao et al., 2020), graph representation learning (Veličković et al., 2018), and medical deep learning (Kaji et al., 2019).

More recently, attention mechanisms have been introduced to the realm of computer vision. These Vision Transformers have demonstrated extraordinary performance and many

works have explored different approaches for integrating them into standard convolutional networks as well as replacing the convolution entirely with just self-attention blocks. Our work builds off of two methods namely Visual Transformers (Wu et al., 2020) and Pyramid Vision Transformer (PVT) (Wang et al., 2022). The first work introduces a semantic grouping module that is able to group pixels in the image and use the pooled features for each group as tokens rather than the singular pixel features. The motivation behind this is that the module creates a more compact set of tokens and thereby reduces the computational cost of the attention module. After obtaining the token-to-token self-attention map, a projection module is used to transform the image features based on the attended group features. The paper integrates the attention module by replacing the last stage of convolutions with their proposed attention architecture, thereby reducing the number of parameters and floating-point operations considerably. PVT approaches the problem differently by proposing to replace convolutions with a custom attention architecture that applies a learned downsampling of the image to a fixed resolution and then attending the image features with the downsampled image features. In this case, the downsampled features preserve global information in the image features. Both models demonstrate improved efficiency and accuracy on ImageNet compared to their convolutional counterparts.

2.1.6 Transfer Learning

Transfer learning is the process by which knowledge accumulated when learning one task can be repurposed to improve baseline performance on another different but similar task. In machine learning applications, this is usually accomplished by training some or all layers of a neural network on one task and then using these trained weights to initialize a model that is then trained for a different task. The process of learning weights from a pretext task is known as "pre-training", and the process of adjusting these pretrained weights on the target task is commonly known as "fine-tuning." Transfer learning can be especially useful in task domains where labeled input data is scarce, including medical imaging contexts.

Datasets used for pretraining models are varied, though use of large-scale general datasets has proven to be effective in jump-starting the training process for many ML applications. In the field of computer vision, the ImageNet dataset from the ILSVRC benchmark is a widely used dataset for performing transfer learning for a number of different target tasks, including image classification, object detection, image segmentation, and action recognition (Huh et al., 2016). The popularity of ImageNet for transfer learning in computer vision is likely due to its large size (1.2 million images) and the large number of object classes (1000). As a consequence of these factors, models pre-trained on ImageNet seem to learn good "general-purpose" features which are amenable to many target task domains. In their experiments with ImageNet pre-training, (Huh et al., 2016) found that reducing the number of ImageNet images by 50% and the number of output classes by over 80% both individually result in only small performance decreases on other image classification benchmarks. They speculate this may mean that pre-training for CNNs is more resilient than previously thought, and that previous estimates of the number and variety

of image samples required to learn good "general-purpose" CNN features were larger than necessary.

More recent works have investigated potential mechanisms for reusing weights learned in 2D classification contexts, such as those from networks pre-trained on ImageNet, to 3D contexts, such as volumetric medical imaging. Some of these mechanisms will be discussed in the following sections.

2.1.7 The I3D Architecture

The Two-Stream Inflated 3D ConvNet (I3D) uses one proposed method for repurposing pre-trained weights from 2D image classification models for 3D imaging tasks (Carreira & Zisserman, 2018). The original paper proposes using 2D pre-trained weights for a spatio-temporal imaging task with video input, though from a theoretical perspective the approach used by I3D can apply to arbitrary volumetric imaging tasks. I3D works by "inflating" 2D $N \times N$ filters into 3D $N \times N \times N$ filters. Initializing these 3D filters is very straightforward: the contents of the 2D filter are simply duplicated at each index along the new axis and normalized by the length of the axis, such that taking their sum along the third axis produces the original 2D filter.

The authors of I3D use this approach to inflate an Inception v1 model with batch normalization, pre-trained on the ImageNet classification task (Ioffe & Szegedy, 2015). They then pre-train the resulting 3D model again on the Kinetics Human Action Video dataset (W. Kay et al., 2017) before testing it on the HMDB-51 (Kuehne et al., n.d.) and UCF-101 (Soomro et al., 2012) action classification datasets. The resulting I3D model achieved state-of-the-art performance on both action classification tasks, highlighting the effectiveness of ImageNet pre-training even for volumetric imaging tasks.

2.1.8 ACS Convolutions

Axial-coronal-sagittal convolutions (ACS) is another proposed method for natively using 2D pre-trained weights in 3D imaging contexts (Yang et al., 2021). The authors of ACS address an inherent flaw in the I3D approach for generalizing 2D pre-trained weights to 3D volumetric imaging contexts. Unlike spatio-temporal applications, where the new temporal axis is semantically distinct from the existing spatial axes, arbitrary volumetric imaging applications do not necessitate a distinction between the new spatial axis and the existing spatial axes. In other words, it is unclear which spatial axis the 2D filter should be oriented to (duplicated across)—in many imaging contexts, including medical imaging, all seem equally valid. The authors address this issue of selecting the optimal axis by avoiding it altogether. Instead of duplicating the 2D filter across a single axis, the $K \times K$ 2D filter is split into 3 separate kernels: one $K \times K \times 1$ filter for the axial axis, one $K \times 1 \times K$ filter for the coronal axis, and another $1 \times K \times K$ filter for the sagittal axis. Due to the threefold increase in the number of filters applied to each region relative to the original network, each of these three resulting filters is normalized by a factor of one third. During an ACS convolution step, these three filters are convolved independently

across the original 3D input. The three outputs are then aggregated either by concatenation or (weighted) averaging, depending on the application.

Like I3D, ACS convolutions provide two major benefits for generalizing 2D convolutional networks to 3D task domains. First, any 2D network architecture can be adapted to 3D input using this scheme with minimal modifications, and second, this scheme permits 2D-to-3D transfer learning by directly loading pre-trained weights from the parent 2D network. In their experiments, Yang et al. (2021) found that a variant of Mask R-CNN adapted to use ACS convolutions consistently benefitted from 2D pre-training on the DeepLesion benchmark, outperforming the previous best approach. Even without pre-training, their ACS model was comparable with or superior to other 3D convolutional approaches that were not as amenable to 2D-to-3D transfer learning.

2.1.9 Deformable Convolutions

While CNNs are well suited for visual recognition tasks they are subject to limitations when they try to accommodate for geometric variations and transformations which can change the position, orientation, and scope of an image. One way to address this is to incorporate these variations in the training dataset. This is typically done by modifying the existing data with affine transformations (Dai et al., 2017). Another method is to apply techniques such as SIFT (scale invariant feature transform) to extract transformation invariant features. However, each method is subject to limitations. The former usually results in expensive training and complex model parameters while the latter might be too difficult to perform for complex variations and transformations. To address these shortcomings Dai et al. introduced deformable convolutions.

Deformable convolutions address the shortcoming of traditional convolutions by adding offsets to the regular sampling grid. Traditional convolutional layers employ a rectangular kernel of fixed size to sample from the input feature map. This introduces a limitation on the layer as it forces all the activation units to have the same receptive field sizes. This makes the model less desirable to use when recognition tasks involve object detection and segmentation. The offsets added to the regular grid by deformable convolutions fix this issue by modifying the constant receptive field of each activation unit. This allows the model to account for data containing variations in image scale.

2.1.10 Multilabel Classification and Triplet Loss

When dealing with numerous classes or a situation where there are only a few samples of data to train on, softmax cross entropy loss likely will not suffice due to the sparsity of the network. One way of combating this issue by employing the triplet loss function (Hoffer & Ailon, 2018). Triplet loss centers around the idea of learning the embeddings of the data in a way such that data points with the same labels have embeddings that are similar to each other and data points with different labels have embeddings that are dissimilar to each other. To achieve this requirement the loss is calculated over a triplet of embeddings: anchor, positive, and negative. The anchor serves as the reference input to which the positive and negative

embeddings are compared to. The positive embedding shares a label with the anchor while the negative embedding label differs from both the anchor and positive. Ideally the distance between the positive and anchor should be small and the distance between and the distance between the negative and anchor should be large. The bulk of the learning occurs when this is not the case and the negative embedding is closer to the anchor causing the model to adjust the weights to reduce the distance to the positive embedding and increase the distance to the negative embedding.

2.1.11 Dropout

Large DNNs that are trained on smaller datasets have a high propensity to overfit the training set due to sampling noise. In theory the best way to address this is to average the predictions from fitting the training set on all possible DNNs. At its full scale this is infeasible but this process can be approximated by using a smaller collection of models called an ensemble. While this process usually improves model performance, averaging predictions of large DNNs is very computationally expensive (Srivastava et al., n.d.). Dropout regularization, proposed by Srivastava et al., addresses this issue by combating overfitting while approximating the ensemble modeling process in an efficient manner.

Dropout works by randomly ignoring certain nodes in a layer. This results in removing all incoming and outgoing connections for that node which effectively removes it from the network temporarily. This results in making the training process slightly more noisy which forces nodes to act more independently as opposed to completely relying on the result of other outputs. This in turn reduces the likelihood of overfitting as the chance of complex relations occurring between the hidden nodes is much less, making the model more generalizable.

2.2 Prior Domain Applications

In recent years, machine learning methods, including deep learning, have seen various applications in psychiatric research. This section will review some areas in which machine learning has been applied to psychiatry, with a particular focus on disorder classification tasks. It is not meant as a comprehensive review, but rather as a demonstration of the variety of methods previously investigated for psychiatric applications and a brief summary of their conclusions.

2.2.1 Alzheimer's Disease

Alzheimer's disease (AD) affects approximately 50 million people globally as of 2020, and is estimated to cause over half of all cases of dementia (Breijyeh & Karaman, 2020). It is currently incurable. As a neurodegenerative disorder, AD is, at least in principle, easier to detect from structural neuroimaging scans than other psychiatric disorders whose impact on brain structure is not as well known. Consequently, many studies investigating techniques for AD classification have relied on structured neuroimaging data as the primary data modality. Organizations such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al.,

2005) have made available relatively large datasets of imaging data from Alzheimer's patients. As a result of these factors, AD has become a significant focus for researchers investigating applications of machine learning to neuroimaging.

Given that neuronal atrophy is a known biomarker of AD, some studies have had success in identifying AD and its precursory states using deep learning-based imaging techniques trained on structural neuroimaging data. Basaia et al. (2018) trained a 3D CNN to distinguish between diagnosed AD patients, patients with mild cognitive impairment preceding an AD diagnosis (c-MCI), and patients with "stable" cognitive impairment that did not lead to AD (s-MCI) from a single cross-sectional sMRI scan. Their trained model was able to correctly distinguish AD patients from healthy controls (HC) with an accuracy of 99% on the ADNI dataset and 98% on an independent dataset assembled using non-ADNI data from two prior studies on AD diagnosis (Albert et al., 2011; McKhann et al., 2011). Cui et al. (2019) take a similar approach to classifying AD and its precursor states using deep learning. However, instead of constructing a CNN from scratch, the authors adapted an existing network, Inception v3 (Szegedy, Vanhoucke, et al., 2015), for this task. Their model was trained and tested on subsets of the ADNI dataset, achieving classification accuracies of 100%, 89.5%, and 77% for AD patients, mild cognitive impairment (MCI), and healthy controls, respectively.

Folego et al. (2018) develop a deep 3D CNN called ADNet for classification of AD, MCI, and HC patients from structural MRI data. Their solution is designed to be entirely automatic, fast, and effective without the incorporation of any prior domain knowledge besides that inherent in the training dataset—no additional clinical information is required besides the scan data, and no brain ROI's are selected prior to training. ADNet is trained on data provided by ADNI and tested on the CADDementia challenge (*CADDementia - Grand Challenge*, n.d.). They also created a second classifier, ADNet with domain adaptation, which achieved a classification accuracy of 52.3% on the CADDementia challenge and was approximately 80 times faster than previous best approaches (Folego et al., 2020).

Importantly, Folego et al. (2018) recognize the importance of ensuring accountability of results obtained by machine learning approaches for diagnostic applications. They consider a few approaches for increasing interpretability of their results, including examination of learned CNN filters, occlusion testing, and visualization of feature representations using t-SNE. Occlusion testing and other methods for improving interpretability will be discussed in chapter 3. They ultimately found that occlusion testing provided the most interpretable information about their model. Using occlusion testing, they found larger activations in their network for AD patient scans in the temporal regions of the brain as well as the posterior cingulate and medial prefrontal cortices, all of which have been previously associated with AD progression.

2.2.2 Major Depressive Disorder

Major depressive disorder (MDD) is a common disorder that has been linked to increased morbidity and mortality across cultures (Kessler & Bromet, 2013). As of 2014, it was ranked by the WHO as the fourth most prevalent cause of disability worldwide. Lifetime prevalence of

MDD across cultures has been estimated between 1.0% (Czech Republic) to 16,9% (United States). Other psychiatric disorders, such as generalized anxiety disorder (GAD) and bipolar disorder (BD), exhibit many overlapping symptoms with MDD, making MDD often difficult to discern accurately in clinical settings (Gao et al., 2018; Hilbert et al., 2017). Consequently, a large portion of the literature on the use of machine learning in automatic classification of MDD has been concerned with distinguishing between MDD and other disorders with similar symptomatic profiles. Unlike AD, less is known about potential biomarkers for MDD, and less studies have been conducted using deep learning techniques and neuroimaging for MDD classification.

Hilbert et al. (2017) take a multimodal approach to distinguish between MDD, GAD, and healthy patients using machine learning. Their results indicated that questionnaire data was useful for case-classification but not disorder-classification, while the opposite was true for imaging data (gray and white matter volumes). Gao et al. (2018) developed an approach to make a similar distinction between MDD and BD using SVMs for classification. They note that, as of 2019, the overwhelming majority of studies investigating machine learning for MDD classification from neuroimaging data use a combination of SVMs or similar approaches and feature reduction steps, rather than deep learning or other methods which operate "natively" on imaging data. Schnyer et al. (2017) also use an SVM to classify MDD patients vs HC patients given a brain map of white matter fractional anisotropy values (FA). Their results supported the hypothesis that relevant information for predicting MDD is not localized, and instead is distributed across many networks within the brain.

Despite the relative lack of deep learning studies for MDD classification from neuroimaging data, some studies using deep learning for MDD classification have been conducted. Mumtaz and Qayyum (2019) developed a deep learning framework to classify MDD based on temporal patterns in EEG signals. Their model was validated on an independent dataset of 63 patients and achieved a classification accuracy of 98.32%, highlighting the potential effectiveness of EEG for MDD diagnosis. Uyulan et al. (2020) also developed a deep learning framework based on EEG for MDD classification. They trained 3 CNNs, ResNet-50 (He et al., 2015), MobileNet (Howard et al., 2017), and Inception v3 (Szegedy, Vanhoucke, et al., 2015), to classify MDD patients and healthy controls. Wang et al. (2021) do use a novel CNN, called 3D-DenseNet, that is designed to operate natively on sMRI. They employ this network with a novel transfer learning approach for MDD, pre-training their CNN on ADNI data for AD classification. Their 3D-DenseNet achieved a classification accuracy of 77.4%. Notably, their ADNI-Transfer pre-training approach improved the classification accuracy of their model by 9.95% compared with training from scratch, suggesting that visual features of neuroimaging data obtained by pre-training on other sMRI classification tasks are highly effective for improving classification performance.

2.2.3 Schizophrenia

Schizophrenia (SCZ) is less prevalent than many other psychiatric disorders, affecting up to 1% of the global population (Javitt, 2014), but it can have debilitating consequences for patient quality of life. Over half of all SCZ patients experience long-term psychiatric consequences, and approximately one fifth of all patients experience chronic symptoms (Owen et al., 2016). Unemployment among SCZ patients is extremely high at approximately 80-90%, and life expectancy is reduced by 10-20 years on average. Like MDD, SCZ is often difficult to reliably identify due to frequent symptom overlap with other psychiatric disorders, especially those which can cause psychotic episodes.

One promising area for identifying potential biomarkers for SCZ is resting-state functional connectivity (FNC) (Arbabshirani et al., 2013; Owen et al., 2016). Arbabshirani et al. (2013) investigate a number of ML approaches for SCZ classification using FNC data. They found that even simple classifiers were able to achieve reasonably high accuracy, e.g., approximately 96% for both k-nearest neighbors (k-NN) and SVM. Kalmady et al. (2019) use functional connectivity (FC) metrics and other regional activity metrics derived from resting-state fMRI to create an ensemble approach for SCZ classification. Their model, called Ensemble algorithm with Multiple Parcellations for Schizophrenia prediction (EMPaSchiz), achieved a classification accuracy of 87% on a dataset that was notably larger than many of those previously used for automatic SCZ classification ($N = 174$).

Like MDD, many existing SCZ classification studies focus on using computationally simpler models with feature reduction steps as opposed to more complex models, such as deep learning approaches, that operate directly on imaging data. However, some studies have been conducted that use deep learning approaches with FC-based features. Zeng et al. (2018) use a deep network called a deep Discriminant Autoencoder Network with Sparsity constraint (DANS) to automatically classify SCZ patients using multi-site functional connectivity MRI data. They train their model on a large dataset of over 1000 participants compiled from prior sources. Their dataset contained a mixture of drug-naive patients and patients who were previously prescribed antipsychotic medications. Their deep network was able to obtain a classification accuracy of 85.0% on a validation set when using pooled data from multiple sites of interest.

2.2.4 Bipolar Disorder

Bipolar disorder (BD) is estimated to affect between 1 and 5% of the global population (Jan et al., 2021)(Jan et al., 2021). BD patients have a high risk of suicide and frequently exhibit self-harm behaviors. Indeed, the life expectancy for BD patients is approximately 9 to 17 years below average (Jan et al., 2021). BD patients may or may not experience psychosis. Given its devastating effects on patient quality of life, early and accurate diagnosis of BD is a high priority for the psychiatric community. BD, like MDD, is frequently prone to misdiagnosis due to its symptom heterogeneity and overlapping symptom profile with other psychiatric disorders such as attention deficit hyperactivity disorder (ADHD), SCZ, and MDD (Anderson et al., 2012) (Jan

et al., 2021). These misdiagnoses can have significant detrimental effects on the quality of patient care and treatment (Jan et al., 2021).

Fung et al. attempt to automatically distinguish between BD, MDD, and HC using an SVM. Their training data consists of two structural neuroanatomical features: cortical thickness and cortical surface area. Their model achieved an overall accuracy of 74.3% (Fung et al., 2015). Like previous SVM-based approaches with similar feature engineering, their approach benefitted from having easily interpretable results. They found that BD patients had greater cortical surface area in the left bankssts, precuneus, precentral, inferior parietal, superior parietal, and right middle temporal gyri compared with MDD patients. Such findings may pave the way for the discovery of potential neuroanatomical biomarkers of BD in the future.

Grotegerd et al. also attempt to develop a framework for automatic discrimination of BD and MDD. An SVM was employed for pattern matching of fMRI scans masked to include regions related to emotional recognition and processing. The model achieved an accuracy of 90% when using fMRI data in happy and neutral states (Grotegerd et al., 2013). While a relatively simple computational model was used for class identification, this research was the first evidence that pattern classification could be used to differentiate unipolar and bipolar depression.

Campese et al. performed psychiatric disorder classification on a BD and SCZ dataset using three methods: SVM, 2D CNN, and 3D CNN. The studies from Fung et al. and Grotegerd et al. show the potential effectiveness of a simple SVM, but Campese et al. proves empirically that 3D CNN models outperform the simpler counterparts. VNet, UNet, and LeNet architectures were compared against classical 2D CNNs and SVMs. VNet achieved the highest average accuracy across two datasets, showing that the effectiveness of 3D convolutional neural networks lends well to neuroimaging and provides a better accuracy than traditional techniques (Campese et al., 2019).

2.2.5 Attention Deficit Hyperactivity Disorder

The pervasiveness of attention deficit hyperactivity disorder (ADHD) is difficult to track. ADHD is a childhood disorder that can continue through adolescence. Studies estimate as low as 1% to as high as 20% of the worldwide school aged population (4-17) has ADHD (Polanczyk et al., 2007). A pooled prevalence puts the more realistic global rates at close to 5% (Polanczyk et al., 2007) in 2007, however that number increased almost 42% between 2003 and 2011 (National Institute of Mental Health [NIMH], 2014). The age of onset for moderate ADHD is six years with children exhibiting difficulties with focus, attention span, controlling behavior, and hyperactivity (NIMH, 2014). Children with ADHD are more prone to troubles with schooling, substance use, and other psychiatric conditions (Evans et al., 2010). With high risks and equally high diagnosis rates, there are also problems associated with overdiagnosis and unnecessary medication. More than two thirds of children diagnosed with ADHD rely on medication as it is the most effective treatment for symptoms of impulsivity, inattention, and hyperactivity (NIMH, 2014) but the long term effects of stimulants in adolescents is not well documented (Ford-Jones,

2015). Proper identification of ADHD will help limit misdiagnosis (as discussed in previous sections), overdiagnosis, and overmedication in these young populations.

Zhang-James et al. used an ensemble classifier consisting of SVM, random forest (RF), k-nearest neighbor (KNN), and gradient boosting (GB) classifiers. 16 principle factors along with age and sex were extracted from the ENIGMA-ADHD dataset of 3,377 structural MRI (sMRI) images. An accuracy of around 66% was achieved. The more important results of this study were findings that intracranial volume, surface area, and subcortical volumes were the most important structures for prediction. The use of an ensemble classifier allowed for interpretable results that help verify our hypothesis of structural biomarkers in the brain, one of which being differences in brain volume.

Mao et al. tackled a harder problem than the previous paper, garnering more accurate results and giving us insight into the benefits of DL and the use of CNNs. The paper proposed a diagnostic method for resting state functional MRI scans (rs-fMRI) which are basically spatio-temporal scans of patients' brains. They used a 4D CNN and tested different methods of granular computing on the ADHD-200 public dataset. An accuracy of 71.3% was achieved without using hand-crafted features like previous papers. While this paper shows the ability of deep learning networks to accurately learn patterns from scratch, its output gives the user less information by virtue of the model selected features.

2.3 Model Implementation v1

As we were collecting data and researching different mental illnesses, we understood that mental illnesses were prone to being misidentified leading to worse care for the patients. Although we identified many datasets that we could use to train machine learning algorithms on, we still faced the question of how to incorporate existing expertise, such as labeled datasets, into these models without being fully reliant on it. On one hand, clinical diagnoses are the only source of ground truth data for detecting mental illnesses from MRI scans, and training a model to correctly predict the disease would make it as accurate as the clinicians' diagnosis. If we were able to gather enough data to train a model that was accurate enough, then it had the potential of working with clinicians to help diagnose patients. On the other hand, as we saw from our research of different mental illnesses, clinicians' biases show up substantially in mental illness diagnosis and using the biased data to train our model will result in a model that reflects the current biased state of diagnosis. Although it is impossible to obtain purely unbiased data, if we wanted to work towards real world application of our model, we would need to at the very least be able to identify these biases as they occurred. Therefore, we decided to use supervised learning with current clinician and researcher labeled data to train our model, and unsupervised learning with the same data to visualize patterns that were unseen by the supervised models. We hope that by combining these two approaches, we could identify these biases and also get high accuracy on our model.

Once we have features from both the supervised and unsupervised models over a lot of data, comparing the two would give us information about the neurophysiological changes that

can be observed for mental illnesses as well as ways to identify if there are biases present in the data. If the supervised and unsupervised models found similar changes in the brain for a particular disorder, then we can say that this disorder has some identifying neurophysiological changes. In addition, we can also obtain more information about the disorders including heterogeneity and symptom overlap. Heterogeneity occurs when people can qualify for a disease with differing symptoms, if our models show that there are varying types of changes that correspond to a particular disease, then we can observe that heterogeneity might be present for that disease. In addition, if the same symptom is present for multiple diseases, and if our model can identify the physiological changes associated with it, we can help identify other factors which can help differentiate the diseases. Lastly, assuming that the models work perfectly, if there is a disagreement between the features from the two approaches, then we can conclude that there are some biases present in the labeled data that should be further analyzed. This can be done by looking at other factors such as race, ethnicity, gender, etc, that can also affect the diagnosis with a professional.

2.3.1 Dataset preprocessing and acquisition

The data that is used for both of our models comes from two main sources. The first source, the same one as used before from the original 3D convolution paper by (Pominova et al., 2019), is the UCLA Consortium for Neuropsychiatric Phenomics. This dataset consists of MRI scans of 272 subjects: 130 healthy subjects, 43 subjects with ADHD, 49 subjects with bipolar, and 50 subjects with schizophrenia (Gorgolewski et al., 2017). We downloaded the dataset from the website and used the dataset in our model as is.

The UCLA dataset was used in our experiments to predict mental illness class from the MRI image. These images were obtained preprocessed, so we did not have to do any preprocessing on the images prior to training our model. The class labels, ADHD, bipolar, schizophrenia, or healthy, corresponding to each scan was provided in a csv file, and were used as ground truth labels while training our models.

2.3.2 Supervised model structure

To start developing a supervised algorithm, we started by using the approach by Pominova et al. (2019). They developed a 3-dimensional deformable convolutional neural network and had trained the model on a dataset available from UCLA Consortium for Neuropsychiatric Phenomics. In 3-dimensional deformable convolutions, there are offsets for each convolution operation which learns to look at a different position than its location. These offsets can help focus the model on the more important parts of the scan and ignore anything that is not useful such as blank space. We wanted to first train the model on the dataset as it was shown in the paper, then use visualization techniques to locate biomarkers in the brain scans. Thus, if a scan was predicted positively by the model, then these visualizations could determine which areas in the brain it looked at to determine its decision. This way we could understand how the model was working, and also uncover patterns in the brain scans which aligned with

each disease. For example, if many of the subjects with schizophrenia had similar changes in the brain, then it could be an indication of a physiological change that can be measured that corresponds with schizophrenia.

2.3.3 Unsupervised model structure

For our unsupervised algorithm, we used an encoder-decoder type autoencoder (AE). Specifically, we decided to use Scale Space Autoencoder (Baur et al., 2020) as shown in Figure 3. First a 2D slice of a brain scan is fed into the model, which performs calculations to represent the scan in a low dimensional space, called the latent dimension. Then, a decoder tries to extrapolate the latent dimensional representation and reconstruct the original scan. Based on the differences of the reconstructed scan and the original scan, the entire model is updated to improve its performance. The Scale Space Autoencoder does this by using a separate autoencoder on different frequency bands of brain MRI as seen in Figure 3. Their method helps improve reconstruction and anomaly detection at higher resolutions (Baur et al., 2020). We used this method for our unsupervised algorithm since our images had larger resolution. As an autoencoder learns, it will learn to represent important attributes in the latent space, meaning if we analyze the latent dimension we should be able to see clusters of features that arise. We wanted to use these features and see if they were similar to the features found in the supervised approach.

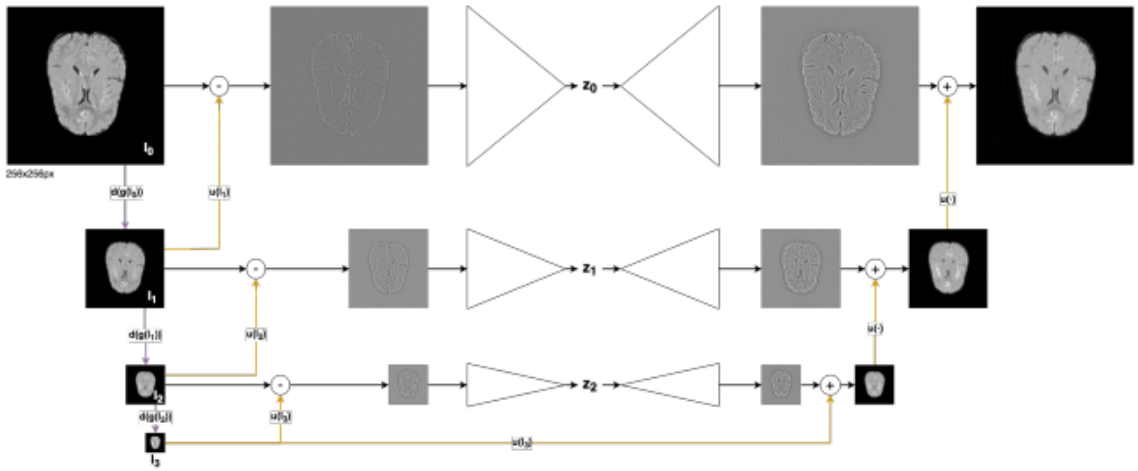

Figure 3. Visualization of different frequency bands used while reconstructing scans. Image taken from Baur et al., 2020.

2.3.4 Supervised Results

For our version of the model we ran two different classification problems. The first was a binary classification experiment between schizophrenia (SCZ) and bipolar (BD) and the second

was a multiclass classification between controls, schizophrenia, bipolar, and ADHD. A version of the multiclass problem was also run with triplet loss. The model performed best on the classification of schizophrenia vs bipolar disorder (light blue) achieving a 59% validation accuracy. For the multiclass classification (orange) problem it achieved a validation accuracy of 46% and 44% for the variation with triplet loss (dark blue). Our validation accuracy curves can be seen in Figure 4.

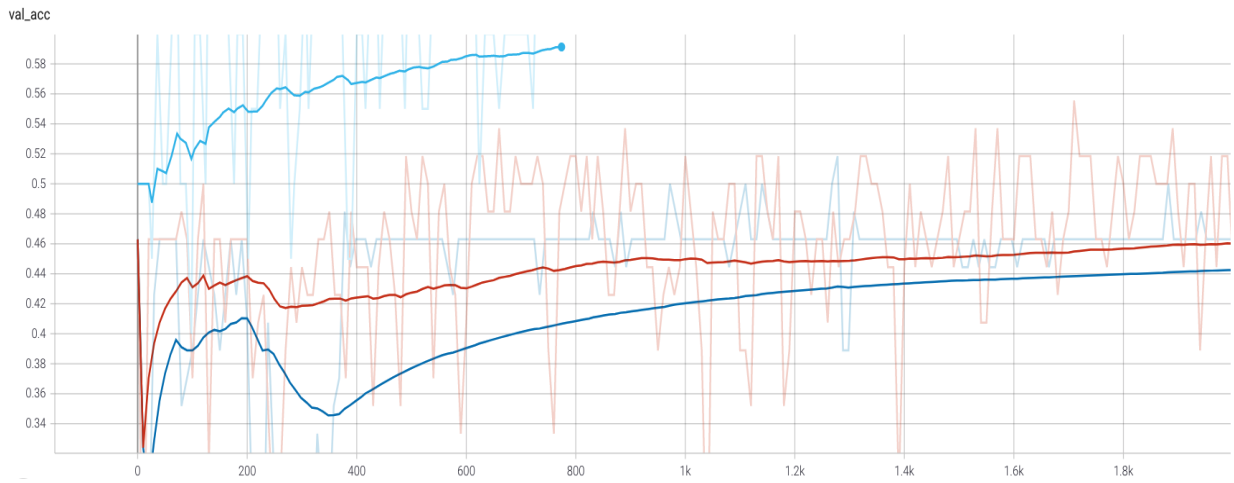

Figure 4. Validation accuracy for the three experiments. Key: schizophrenia vs bipolar disorder (light blue), multiclass classification (orange), and variation with triplet loss (dark blue).

A key component of our model was the visualizations that could be generated using the deformable offsets (see Section 3.1.3). With offsets it is possible to highlight potential regions of interest in the image by clustering the voxels with greatest deformation. In Figure 5 below, the red dots indicate the voxels of greatest deformation the model sampled when predicting diagnosis. The colored polygons indicate regions of points that were closer together than other points. These areas potentially indicate regions of greatest interest to the model.

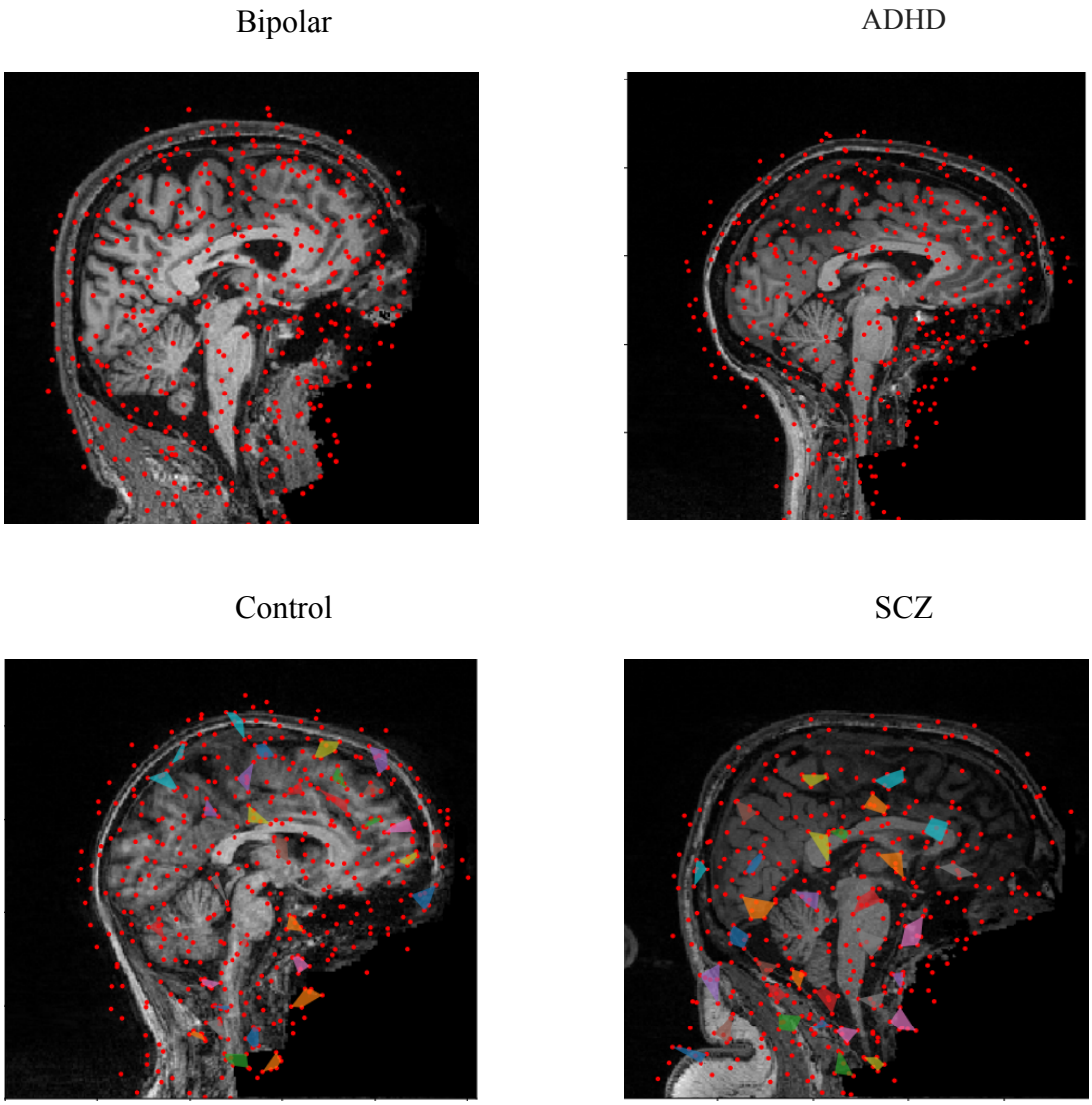

Figure 5. Visualization of offsets with greatest deformation. Red dots represent areas of higher clustering of the deformable offsets in the model. Shaded polygons represent areas that show exceptionally high levels of clustering. Clustering represents areas that the model looked at while deciding its prediction on the scan's phenotype.

2.3.5 Unsupervised Results

We adapted the architecture for the Scale Space VAE and attempted to apply it for reconstructing patients in the control section. The motivation of the architecture was to increase

focus on the high frequency features as classic VAE architectures generally have blurry outputs, with the assumption that these high frequency features contain the discriminating information between MRI scans from different disorders. However, we found that it was much more difficult for the VAE to reconstruct the high-frequency features as can be seen by the blurry reconstructions in the left figure, as shown in Figure 6. The analysis of the frequency features on the right demonstrate that while some of the high frequency information such as the brain boundaries and some lobe boundaries are preserved, most of it is lost. Due to additional time constraints, we were unable to further improve on this architecture.

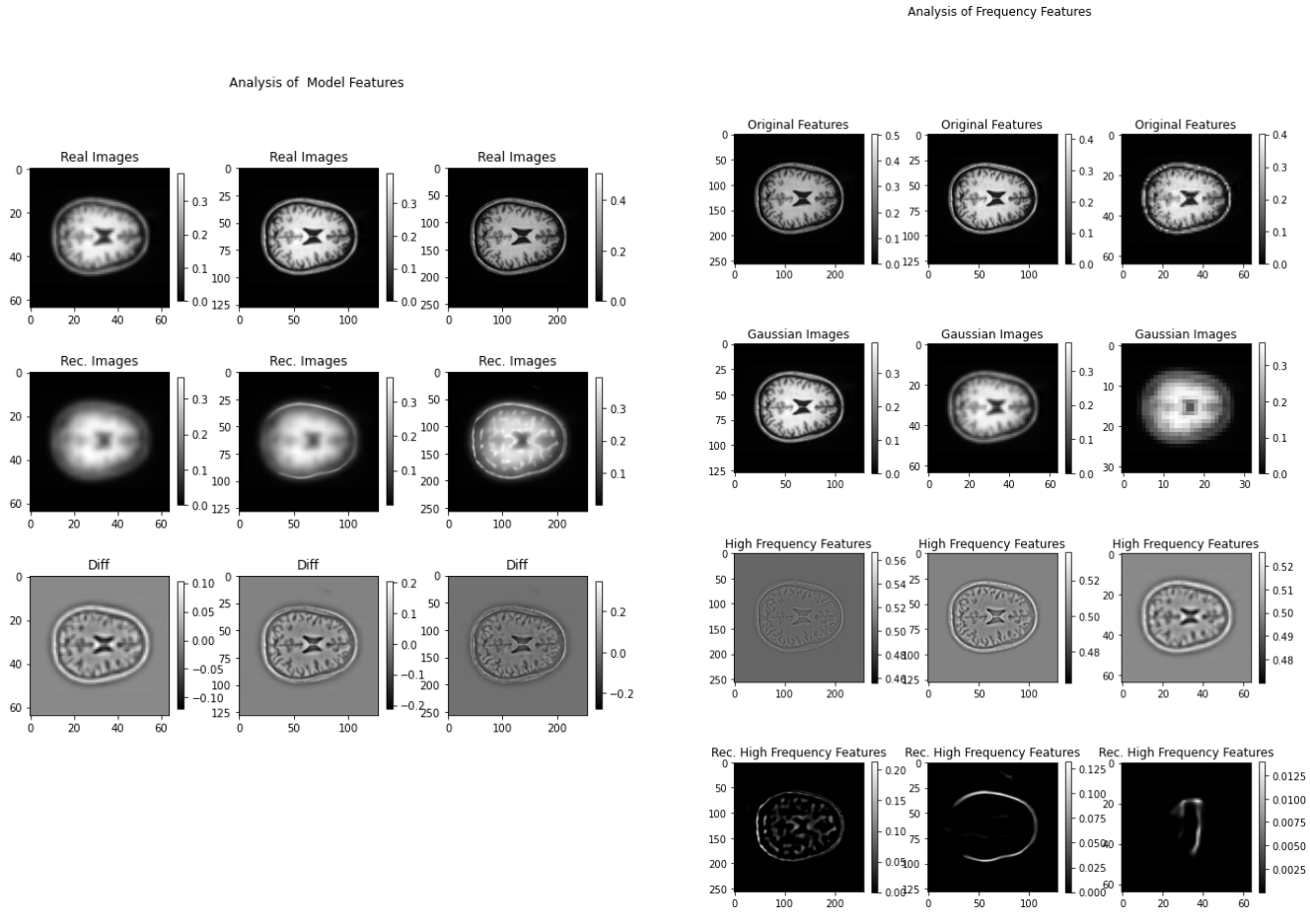

Figure 6. Unsupervised model outputs. Model images are shown along with reconstructed (Rec) images. High frequency images show the detail obtained from higher frequencies from analyzing outputs from each level in the model.

2.3.6 Lessons Learned

Offset visualizations for the deformable vox resnet from version two can be seen in the following figures. The deformable vox resnet was trained on two datasets UCLA and NDA. In the prior case a binary classification was run between control vs presence of a mental illness and in the latter a multi label classification was performed on three symptom labels.

The supervised model was on team members' computers using their GPU resources. We saw that even if we followed the same hyperparameters used by Pominova et al., 2019 we were not able to replicate the general performance of the 3D deformable network. It should be noted that our classification problems differed slightly from those performed in (Pominova et al., 2019). As opposed to a dichotomy between diseases and controls, our experiments primarily focused on differentiating between two diseases and multi class experiments involving classification between three diseases and healthy controls. Of the different experiments we ran the schizophrenia vs bipolar classification performed the best which matched our initial predictions based on the performance of the 3D deformable net from Pominova et al. on classifications of schizophrenia vs healthy controls and bipolar vs healthy controls as these had accuracies of 82% and 68% respectively.

For the unsupervised method, we were unable to train the model well enough to get meaningful results. The model was very large and had unstable loss values that did not decrease. In addition, unsupervised methods often require large amounts of data to train properly and just using the UCLA Consortium for Neuropsychiatric Phenomics dataset was not enough. It should be noted that our classification problems differed slightly from those performed in (Pominova et al., 2019). As opposed to a dichotomy between diseases and controls, our experiments primarily focused on differentiating between multiple diseases with the multi class experiment involving controls.

2.4 Model Implementation v2

Since our main challenge in our earlier experiments was that we did not have enough data to train a robust machine learning model, we wanted to gather more data. In addition, while dealing with these issues, we thought of an alternative using the patient symptoms as the truth values instead of diseases. Given our end goal of trying to visualize the irregularities in the brain we thought that training the model on patient symptoms would give a better chance of potentially mapping the symptoms to said irregularities. Potentially this model could be used to measure symptom intensity as well and also help us with our inquiry into the symptom overlap problem between multiple diseases.

2.4.1 Dataset preprocessing and acquisition

As before, we used the dataset from Gorgolewski et al. (2017) in our experiments to predict mental illness class from the MRI image. The class labels, ADHD, bipolar,

schizophrenia, or healthy, corresponding to each scan was provided in a csv file, and were used as ground truth labels while training our models.

In addition to the UCLA dataset, we obtained access to the NDA/NIMH data archive which gave us a large amount of datasets to use. We wanted to first focus on psychosis for our analysis to start and so we downloaded two datasets that had a large amount of MRI images (Tamminga, 2017, 2020). Both of the NDA/NIMH datasets were not preprocessed when we obtained it. We developed a preprocessing pipeline which took as input the NDA downloaded file directory and outputted a preprocessed dataset which could be used in our model. These datasets included the PANSS scale which measures many symptoms that are present in mental illnesses (S. R. Kay et al., 1987).

Our preprocessing pipeline consisted of several steps to filter, convert, and move MRI scans. First, we filter the image description table using the regular expression r'.*MPRAGE.*|.*mprage.*|.*T1.*', in order to obtain only the MRI scans that we are interested in. Then, we remove duplicate values from the PANSS diagnostic table, the image description table, and the subject description table so that we only have one entry per participant. Then, we join the tables together to create a large table and unzip the images that we are interested in and convert them from dicom format to niftii format. We use the pydeface tool to deface the images so that the identifiable parts of the image, such as facial features are censored when training the model (Gulban et al., 2019). This is to reduce the chance that the model learns something that is not important and reduce bias in the model's predictions. Once we have our dataset, we create a yaml file and use it for training our model. Our model code is developed using the TorchIO framework which makes data processing medical images for deep learning models using pytorch efficient and easy to use (Pérez-García et al., 2021). Our model code is available online on Github.

Access to an NDA dataset helped us realize some of these ideas as it gave us a much larger sample size which translates into a more robust and less biased model. Specifically, we obtained 291 images from the first NDA study we downloaded and 24 from the second and decided to use the former for our experiments since it had a substantial amount of images. The dataset also came with symptom information in the form of PANSS scores which are based on interviews conducted with the patient. PANSS scores, which would act as the regression values for the model, record the intensity of different types of symptoms making it very useful for capturing symptom intensity in our visualizations which was one of our goals when incorporating symptom based classification. We used the column from the PANSS questionnaire which rated the severity of the symptom of interest from 1 to 7, with 1-4 being low, and 5-7 being high.

2.4.2 Atlas Attention Module

This section describes the motivation behind our atlas attention module and its architecture. Inspired by Wu et al. (2020) and Wang et al. (2022), we aim to group the MRI scan into groups of meaningful components. Rather than using a learnable module which would

require more data, we utilize an off-the-shelf segmentation model that is able to generate roughly accurate atlases for a scan. We then use the segmentations to pool the image features. The segmentation model can partition the scan into at most 95 different segments. We then add a separate segment to distinguish between empty regions within the brain and the background. We use these pooled features to generate our region-wise attention map which is able to capture global feature information across each slice. After attending the global features, we then project them back to the image space using the aforementioned approach. This mechanism is closely related to the PVT attention module except rather than using a downsampled feature to capture global features, we use the atlas regions. The benefit of this is that we can better identify meaningful regions that correspond to the potential presence of a specific mental illness. Since each atlas is slice dependent, we ensure that our base model does not pool in the slice dimension. In order to facilitate slice-wise dependencies though, we add a temporal convolution prior to applying the atlas attention module.

2.4.3 Supervised model structure

Unfortunately, during our exploration of unsupervised methods for identifying latent classes in our MRI data, we ran into a number of computational challenges. Despite many efforts to optimize hyperparameters, the performance of our unsupervised models was unable to compare with that of our supervised learning experiments. These reasons have been mentioned earlier but an additional explanation for poor performance could be the smaller size of our training dataset(s). While the unsupervised methods might theoretically avoid the pitfalls of biases in diagnostic labels, since they have no prior knowledge to rely on, deconstructing the input data generally requires more samples for results to be practically useful. Considering both the large size and high-dimensionality of individual MRI input samples and the relative scarcity of large MRI datasets, we simply did not have a large enough training set to use unsupervised learning approaches effectively. With these two limitations in mind, we eventually decided to shift our attention away from the unsupervised methods and focus entirely on our supervised models.

In doing so, we recognized that our continued reliance on predetermined diagnostic labels meant that the results of our supervised approach would be subject to the same criticisms as existing symptom-based classification, such as heterogeneity and over-/under-specificity. In an attempt to remedy this, we shifted away from single-label classification using diagnoses and adopted multilabel classification using symptom-based dimensions. These symptom-based labels were obtained by the method described in section 2.4.1. Specifically, we used the columns for delusions, hallucinations, and anxiety for our experiments.

After addressing issues with compute resource requirements and reliance on existing diagnostic labels, we wanted to improve the interpretability of our model's results. Interpretability presents a major obstacle to the implementation of medical artificial intelligence in clinical practice. One reason for this is that it is often difficult to understand the output of deep learning models in medicine within the context of existing domain knowledge. Explaining

reasons which led to a specific diagnosis are important for building trust and understanding between patients and medical professionals. Consequently, models that help bridge the gap between "black-box" predictions and the domain knowledge of human practitioners will likely have a much better chance of being effectively deployed in a clinical setting. To this end, we sought to incorporate some information about structural and functional neuroanatomy into our model via a novel Atlas Attention module.

2.4.4 Ablation Models

To add rigor to our experiments, we compared the performance of our Atlas Attention module with other imaging models for MRI vision tasks. These models included a deformable VoxResNet and 2D-to-3D models such as I3D and ACS. In all cases, these models were adapted to our task by adjusting input and output dimensions and, in the case of the I3D and ACS models, initializing weights from a ResNet model pre-trained on the ImageNet object classification task. These models were compared in multiple experiments on both the UCLA and PANSS datasets. For the Atlas Attention network, we experimented with various layer orders to manipulate the relative location of the Atlas Attention module. We also experimented with various hyperparameters such as learning rate and learning rate decay for all models to try and optimize our model training.

2.4.5 Results

ABLATION - CLASSIFICATION - UCLA - Control vs {Bipolar, Schizophrenia, ADHD}					
dVoxResNet	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size**

					(MB)
Vanilla	69.34%	68.83%	2,545,810	1.00E-05	2,562.98
I3D - W/O ATLAS	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Inflated	66.03%	66.03%	33,167,298	1.00E-05	21,310.53
Not Inflated	66.03%	66.03%	11,177,538	1.00E-05	21,226.64
I3D - w/ Atlas	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Inflated			15,474,114	1.00E-04	13,619.07
Not Inflated	66.04%	65.93%	9,999,426	1.00E-04	13,597.17
ACS Convolution	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Vanilla	62.26%	60.57%	11,177,538	1.00E-05	31,810.64

Table 1. Results on UCLA dataset while training control vs. all disorders (bipolar, schizophrenia, ADHD).

ABLATION - CLASSIFICATION - UCLA - Control vs Bipolar					
dVoxResNet	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Vanilla	71.43%	41.67%	2,545,810	1.00E-05	2,562.98
I3D - W/O ATLAS	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Inflated	77.14%	59.77%	33,167,298	5.00E-05	21,310.53
Not Inflated	82.86%	73.21%	11,177,538	5.00E-05	21,226.64
I3D - W/ ATLAS	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Inflated			15,474,114	5.00E-05	13,619.07
Not Inflated	71.43%	41.67%	9,999,426	5.00E-05	13,597.17
ACS Convolution	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Vanilla	Skip	Skip	Skip	Skip	Skip

Table 2. Results on UCLA dataset while training control vs. bipolar.

ABLATION - CLASSIFICATION - UCLA - Control vs Schz					
dVoxResNet	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**

Vanilla	85.71%	83.81%	2,545,810	1.00E-05	2,562.98
I3D - W/O ATLAS	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Inflated	88.57%	85.04%	33,167,298	1.00E-05	21,310.53
Not Inflated	85.71%	83.81%	11,177,538	1.00E-05	21,226.64
I3D - W/ ATLAS	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Inflated				5.00E-05	13,619.07
Not Inflated	91.43%	88.35%	9,999,426	5.00E-05	13,597.17
ACS Convolution	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size**
Vanilla	Skip	Skip	Skip	Skip	Skip

Table 3. Results on UCLA dataset while training control vs. schizophrenia.

ABLATION - CLASSIFICATION - PANSS					
dVoxResNet	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Vanilla	50.84%	86.27%	2,545,939	1.00E-03	2,562.98
I3D - FOR COMPARISON W/ ATLAS	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Inflated - No ATLAS	37.28%	71.73%	33,167,811	1.00E-03	21,310.53
Not Inflated - NO ATLAS	44.06%	82.86%	11,178,051	1.00E-03	21,226.64
Not Inflated - ATLAS	Skip	Skip	Skip	Skip	Skip
Inflated - ATLAS	Skip	Skip	Skip	Skip	Skip
ACS Convolution	**Accuracy**	**F1**	**N_params**	**LR**	**Model Size (MB)**
Vanilla	Skip	Skip	Skip	Skip	Skip

Table 4. Results on NDA dataset while training on three symptoms.

2.5 Discussion

Our experiments with the UCLA dataset consisted of control vs. multiple disorders(Table 1), control vs. bipolar disorder (Table 2), and control vs. schizophrenia (Table 3). For each of

these experiments, a random chance would have an accuracy of 50%. All of our models that we trained performed better than random chance. For our experiments with the NDA dataset (Table 4), the model was classifying if symptoms in interest would be classified as high or low. In these experiments, three symptoms were used: delusions, hallucinations, and anxiety. Random chance accuracy would be at 12.8%. Some experiments had to be skipped due to memory limitations and time constraints during training.

For the control vs. ill experiment, the deformable dVoxResNet performed the best, for the control vs. bipolar experiment, the non-inflated I3D had the best performance, and for the control vs. schizophrenia experiment, the inflated I3D performed the best. For the experiments with PANSS, the deformable VoxResNet also performed the best. Our learning rate was set at 1e-5 for the UCLA experiments with a 1/10 decay every 5 epochs. For the NDA/PANSS models, our learning rate was set at 1e-3 and a decay of 1/10 every 15 epochs. For both datasets, we trained our models for 50 epochs.

For the UCLA dataset, healthy vs. schizophrenia (Table 3) classification had higher performance than the healthy vs. ill (Table 1) and healthy vs. bipolar disorder (Table 2). This was consistent with results seen in the original deformable offset method, which also had the highest performance for the healthy vs. schizophrenia classification (Pominova et al., 2019). In addition, the atlas attention model integrated with the non-inflated I3D was able to improve the performance for health vs schizophrenia on the UCLA dataset to 88% F1 and 91% accuracy compared to existing models. However, the model performed slightly worse for the overall healthy vs ill classification compared to I3D and on par with dVoxResNet on the healthy vs. bipolar at 41.7% F1. This could be due to convergence issues since the F1 and accuracy for I3D without inflation was around 73% and 82% respectively. Furthermore, adding atlas attention as the last stage led to a significant decrease in number of parameters and model size.

Interestingly, the model trained on NDA was able to learn and differentiate symptoms fairly well. As seen in Table 4, the best NDA model was trained using the deformable VoxResNet and got an F1 score of 94.64% for predicting delusions, 67.46% for predicting hallucinations and 97.39% for predicting anxiety and an accuracy of 50.84% overall. This seems to show that it is possible to predict these symptoms from MRI scans.

While these results show that the model was able to recognize some changes in the brain that helped differentiate the diseases, they are not very useful without visualizations to help see which regions are affected by the diseases.

3 Explainable Methods and Design

Explainability and interpretability within AI are the key issues in trust critical applications. The machine learning algorithms that give the best performance are often seen as black boxes and thus considered unreliable. Fundamentally, all methods of model interpretability work to build a pipeline for unveiling transforms that map inputs to outputs. This is facilitated by showing how the model came to its conclusion in a way that is interpretable to a user (Chromik et al. , 2021). For ML algorithms to enter clinical use, the results of a model must be presented in a way such that a non-expert can understand and reason with its conclusions. Counter examples, perturbations, and varied levels of explanation are necessary to verify a model is returning explainable results (Zhang et al., 2022; Chromik et al., 2021). Thus, it is crucial to make the model as transparent as possible to allow for users to examine its predictions.

3.1 Foundational Literature

3.1.1 High level ML Taxonomy

Explainability comes in two main variants, transparent models and post-hoc explanations. Transparent models are defined by 3 tenants: "simulatability, decomposability and algorithmic transparency" (Barredo Arrieta et al. 2020). Simulatablity implies that a human can reason through how a model came to its results. Decomposability is the ability to explain the model piece by piece in a manner humans can understand. Algorithm transparency is the principle that mathematical analysis and related methods, which may be hard to conceptualize, can explain model function (Lipton n.d.).

Post-hoc methods, rather than explain the exact function of the model via its structure, attempt to explain model function via forms of metaphor or model attention extraction. For example, local explanations clarify inner function by demonstrating the effect of a single feature or element on the output (Barredo Arrieta et al. 2020). The most useful explainability methods are forms of visual explanations which try to illustrate model behavior. The following sections show implementations of various post-hoc explainability methods.

3.1.2 Deep Neural Network Explainability Methods

DL models are largely interpreted through forms of post-hoc explanations. Methods of explanation for DNN's analyze the network gradients or propagate through the network to unearth behavior. Sensitivity analysis and Taylor decomposition rely on gradient analysis while deconvolution, guided backpropagation, and layer-wise relevance propagation propagate through the network. Sensitivity analysis measures activation gradients to determine the relevance of certain features in a sample. Taylor decomposition attempts to create a relevance metric that combines the local sensitivity at a point and the impact of that value on a particular prediction.

This method tends to yield more complete results than sensitivity analysis alone at the cost of magnifying negative relevance (Montavon et al., 2018).

The following methods are forms of backpropagation which use the DNN structure to propagate and then pool activation in reverse. Deconvolution creates a parallel structure similar to that of the original model. Features at various layers are marked and used to reconstruct the DNN's feature outputs via unpooling (Zeiler et al., 2014). Guided backpropagation functions by using a deconvolution and backpropagation together. It takes values coming from the features unpooled in the deconvolution and the values returned from the backpropagation and when negative masks it. This leads to higher fidelity results (Springenberg et al., 2014). Layer-wise relevance propagation uses a forward pass where activations are collected, a computed score from network output is then back propagated via a set of rules which change the relevance valuation per neuron. This method tends to have more flexible visualizations and can function after transfer learning (Montavon et al., 2018).

3.1.3 Deformable Offsets

While traditional CNN networks have performed well in various CV tasks, they are inherently limited by the regular shape of the convolution kernel. Deformable convolutions allow for the kernel points (ie. the location of the convolution) to be learned as well. These allow for the model to pick up on more geometrically complex input data (Dai et al., 2017). This approach has proven useful in medical imaging analysis as seen in Pominova et al. 2019. Deformable CNNs were more accurate than comparable CNN variants in classification of both processed and unprocessed static MRI images. Additionally, deformations have proven useful for medical segmentation tasks. Li et al. demonstrated the networks were able to pinpoint more clear and anatomically specific regions of interest in high noise tumor segmentation. Deformations also have shown promise in creating explainable part models. The principle behind this approach by Donnelly et al. 2021 is the model can be trained to find significant features, and when shown another member of the same class deform the 'prototype' for that class onto the other image as a way to examine similarity. Finally, Zhu et al. 2018 introduced the principle of added modulation to the convolutional deformation which incorporates spatial attention metrics that effectively focus the deformation by zeroing out areas found to be irrelevant. This proved to be more accurate and create more clear regions of interest mappings. Deformable convolutions embed the model with region-of-interest mappings that can be recovered from the learned defformed offsets. This creates a post-hoc but non-model agnostic method to display attention visually.

3.1.4 Saliency Visualization

Saliency maps display class appearance models via a strategy of backpropagation and maximization per class (Simonyan et al, 2014). Class appearance models are how an image class looks to the model. This directly leads into back propagation based methods that take advantage of gradient propagation for an input as discussed in Montavon et al. (2018). One such approach is Grad-CAM by Selvaraju et al. (2020) which utilizes gradient information from convolutional

layers to assign importance to neurons. Backpropagation then allows pixel importance to be determined after each layer gets average pooled. These methods are model-agnostic, assuming convolutional layers are present and have been implemented in a flexible framework as in Gotkowski et al. (2020). This enables quick application of Grad-CAM and other gradient activation based saliency map methods. In addition, there are other methods of saliency that go beyond back propagation based methods. Mundhenk et al. (2020) proposes a more efficient saliency map generation method. It takes the output at each scale layer, the layers before downsampling and filtering, uses them to measure that activations by an input and then computes the informativeness of the output activation. After being done for each scale they are combined together with a threshold of saliency in a particular color coding scheme that indicates which scale layer the pixel coloring originates.

3.1.5 Text Based Explanation

An alternative to visual explanations for DNN models is the use of textual explanation for key features in image classification. Kim et al. (2018) proposes a Concept Activation Vector (CAV). A CAV creates mappings from set concepts an image could hold and then maps samples into hyper planes that may or may not match. Directional derivatives help gauge the sensitivity of a region of perturbation in relation to how it corresponds with a concept. Hernandez et al. (2022) proposes a similar but more unsupervised method Mutual-Information-Guided Linguistic Annotation of Neurons (MILAN). Activation masks are computed for the input images. A probabilistic estimation of what a human would describe a region and or one neuron is made. Image regions probabilities are approximated by a Show-And-Tell model (Xu et al. 2016). While the probability that any neuron would be described in a manner is approximated with a two-layer LSTM on the annotations language of MILAN (Hochreiter et al. 1997). The estimations are derived from a human annotation of known image sets like ImageNet and Places365 for specific local information rather than full scene descriptions. Then the probabilities of local area description and overall description are weighted and a selection of highly probable terms are selected via a beam search.

3.1.6 Quantitative Evaluation

Visual explanation methods can be evaluated by robustness, fidelity, and contrastivity. Robustness is a measure developed by Alvarez-Melis et al. 2018 which follows the implicit idea that when an input image is minorly perturbed it should still receive a similar visual attribution. This was measured as a function of local stability with local Lipschitz continuity, measured with Local Interpretable Model-Agnostic Explanations (LIME) by Ribeiro et al. (2016) and Shapel Additive Explanations (SHAP) by Lundberg et al. (2017). Fidelity was defined by Pope et al. (2019) as the change in accuracy related to occlusion of pixels with high saliency. Pixels with sufficient saliency are directly related classification and occluding them could lead to changes in accuracy. A similar method of evaluation was proposed by Mundhenk et al. (2020) in the Keep And Retain (KAR) and Remove And Retain (ROAR) methods. The prior keeps high saliency

information under the proposition that it should be the only important information and the latter keeps the least salient information under the influence that accuracy should drop sharply if the saliency corresponds to actual model function. The performance of the saliency method can then be seen via accuracy comparisons when KAR and ROAR are taken at the same threshold of saliency among different explainability methods. Contrastivity defined as by Pope et al. (2019) is a measure of the uniqueness of the visual mappings each class received. This is computed by taking the hamming distance of the binarized saliency maps attributed to different classes.

3.2 Implementation

When implementing explainability methods into our MRI classification models we had two key issues: 1) How can we extract model attention in a way that does not impact performance? 2) How can we ensure these features would be relevant to a clinician?

Our first approach was a non-model-agnostic deformable offset image annotation method. It used the underlying attention metrics computed by the deformable offsets to reveal the significance of certain areas of the brain (Zhu et al. 2018). Our second involved the use of Grad-CAM techniques to extract backpropagation saliency to display model attention in another format. This latter method was non-model-agnostic which proved helpful in later versions of ablation testing and model design reconfiguration. To help increase the relevance of saliency and offset annotations we found that parallel segmentation of input MRI images allowed for rapid overlay of known functional regions of the brain onto our predicted regions of interest. Finally, we integrated a segmentation based occlusion function which would use saliency metrics combined with the known functional regions. In areas with large overlap, we tested the robustness and other quantitative metrics to measure how the model changed in performance.

3.2.1 Deformable Offset Display

Within the first model structure modeled based off of Pominova et al. (2019), the activations of the offsets for the deformable convolution could be extracted by a PyTorch hook. Displaying the offsets made it clear that filtering was needed, otherwise the input MRI becomes imperceptible due to too many dot annotations. The process to filter and then display the dots was as follows: measure offset displacement within the slice, select kernel representative point, select kernel points based on a threshold and modifier, rescale and map kernel points to original input scale, dot information encoding, and dot grouping. Displacement measurements were taken by computing the per kernel point offsets in all dimensions and then computing it into a euclidean displacement vector.

Kernel point representatives were selected via a kernel point selector. Two selectors were used, the mean point and the extreme point. The mean point takes all displacement vectors in a kernel and produces the average vector. Given that all points in a kernel exist within the same local region, they should have similar displacement and grouping them should more coherently group information. The extreme point takes the kernel points and selects the absolute largest one

as the representative. The intuition being the most extreme movement is linked to the area of highest attention and should be selected.

Once representatives are produced for all vectors they are then set for annotation by checking if the length of the displacement vector for that point is over a target threshold. The higher the threshold the fewer points there are in the annotation. This leads to more sparse groupings. The threshold was a scaled mean of the lengths of all the representative vectors. After the final set of points was selected, they are geometrically rescalled to match the initial input dimension. The size difference between the initial input and the deformable convolutional layer input are commuted to perform the rescaling. At times, due to the selection method, total point wide deviations could be seen so a flexible shifting constraint was added to recenter points.

Points were then displayed in a MatplotLib annotation of a point scatter over the original image slice. These points had their size and transparency scaled to be less than one image cell to preserve visibility. Their relative hue, which were shades of red, was determined by either the relative offset in the dimension that would point outward from the 2D view of the 3D scan or they were set to all be equal. The points also had applications of other annotations like quivers applied to show the relative displacement between points. This helps point to areas of displacement convergence in the image.

Finally, offset points were clustered into groups to illustrate areas of specific attention. These were done via K-Means clustering and other similar algorithms like OPTICS (Likas et al., 2003, Ankerst et al., 1999). These clusters were then grouped to make a convex hull to identify a larger area of significance around the relevant point.

3.2.2 Grad-CAM Saliency Maps

As the model structure changed and limits of the deformable offsets became apparent, an alternative model attention display mechanism was necessary. Grad-CAM was implemented via the existing M3d-CAM toolbox from Gotkowski et al. (2020). Saliency maps for each layer were analyzed; earlier layers gave the most directly relevant features to the input image and thus were used as the main saliency visualization. However, the individual fidelity of any saliency map was low. In general, most of the MRI received some form of attention. To focus in on the most critical features, the saliency score for each voxel was analyzed and removed if it fell below certain thresholds. The thresholds were set at one standard deviation above the mean. This augmentation was joined by layered saliency views to encode multiple layers onto the same view as done in Mundhenk et al. (2020).

3.2.3 Atlas Segmentation

Brain atlas segmentation was performed with the same code base as the atlas attention module. For visualization and comparison purposes, the quick segmentation network of Henschel et al. (2022) was used. This leverages the pre-trained FastSurferCNN network to parse the MRI image into 95 distinct brain structures. These distinct structure segmentations are then

taken as voxels points and matched with existing dot annotations and Grad-CAM derived saliency maps. Following this, areas of individual focus are identified by brain region label.

3.2.4 Occlusion Methods

Occlusion methods were incorporated in similar style to Alvarez-Melis et al. 2018 and Mundhenk et al. (2020) where areas of high / low saliency were defined by Grad-CAM based attention identification. Effectively these zeroed out sections of the image as a perturbation to check the robustness of the explainable visualization based on what was identified with full information. Additionally these were also targeted based on brain regions that reached high saliency to see the effect of total region elimination on model accuracy.

3.3 Results

Initial uses of explainability visualization proved useful for fast model sensibility checking, as when model attention cannot be investigated large errors can be introduced and remain undiscovered. Earlier model training operations required the use of the Pillow Python library. This is one of the commonly used libraries for image processing and data pipelines for machine learning. However, due to the transformations the MR files would take when subjected to library interpolation it effectively inverted the image. This led to the model moving its attention away from the actual point of interest, the brain and to the irrelevant empty space around the patient's head in the MRI scan.

Visualizations with offset dots in the first model version can be seen below. There are clear groupings of attentive points around the brain region and specific clusters as shown in Figures 8-10, with offset clustering which could point to brain areas of importance. Generally, for images from the UCLA datasets, Figures 7-9, they cluster around the frontal cortex in all cases than Schizophrenia. This indicates that areas of the brain linked with higher functioning may display a correlational relationship with certain mental illness designations. For images generated from the NDA dataset, the clusterings are less clear and harder to characterize (figures 10-14). Which likely indicates that the model had difficulty focusing on specific areas in particular for the multi label classification. Further analysis of the specific clustering in relation to brain segmentations and metrics on the sparsity of these clusterings should be done in the future.

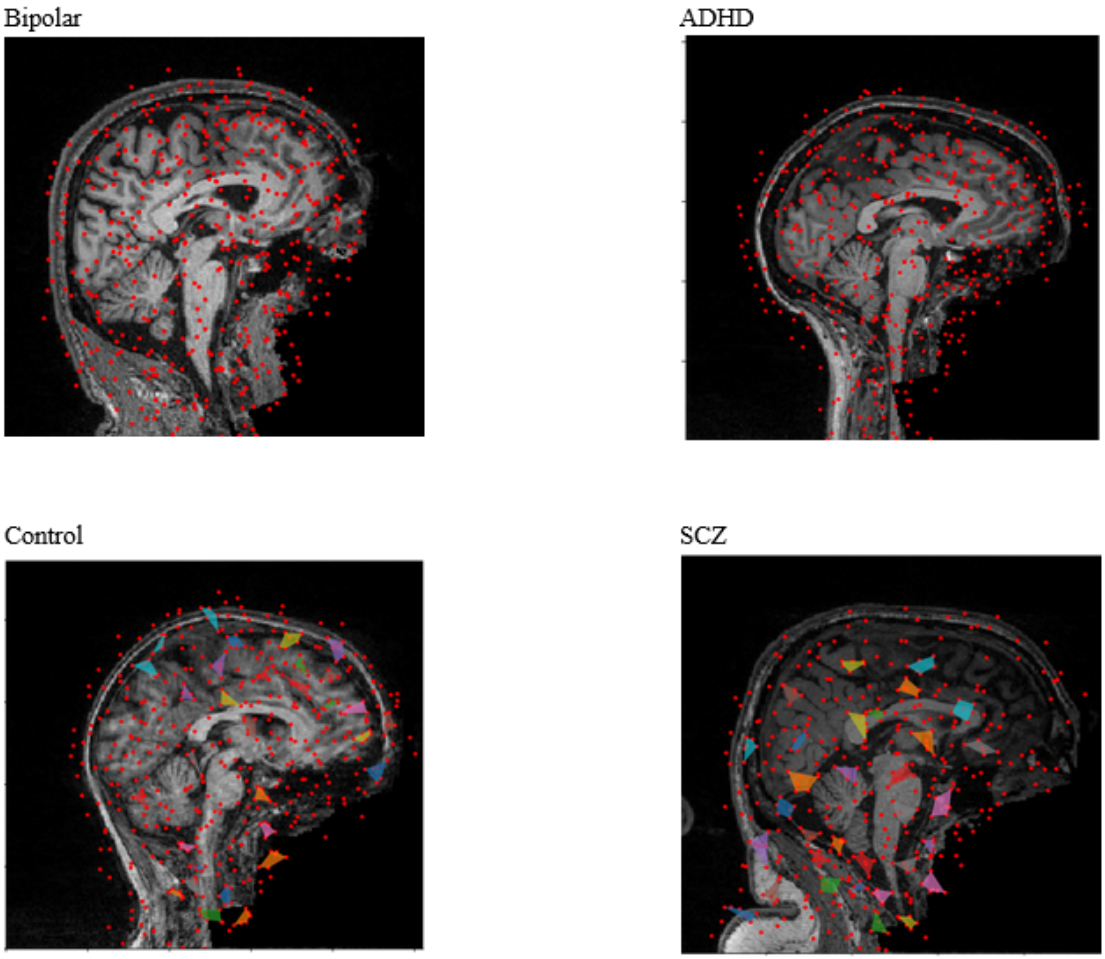

Figure 7. Dot offset display Clustering for UCLA dataset in model version 1. Clustering from dVoxResNet models trained on all four classes. Offset visualization for bipolar, ADHD, schizophrenia and control patients. Dots indicate areas of offset clustering and areas that the model looked at to decide diagnosis. Colors of regional areas indicate close clusters of points, the color is to distinguish groups not for image segmentation.

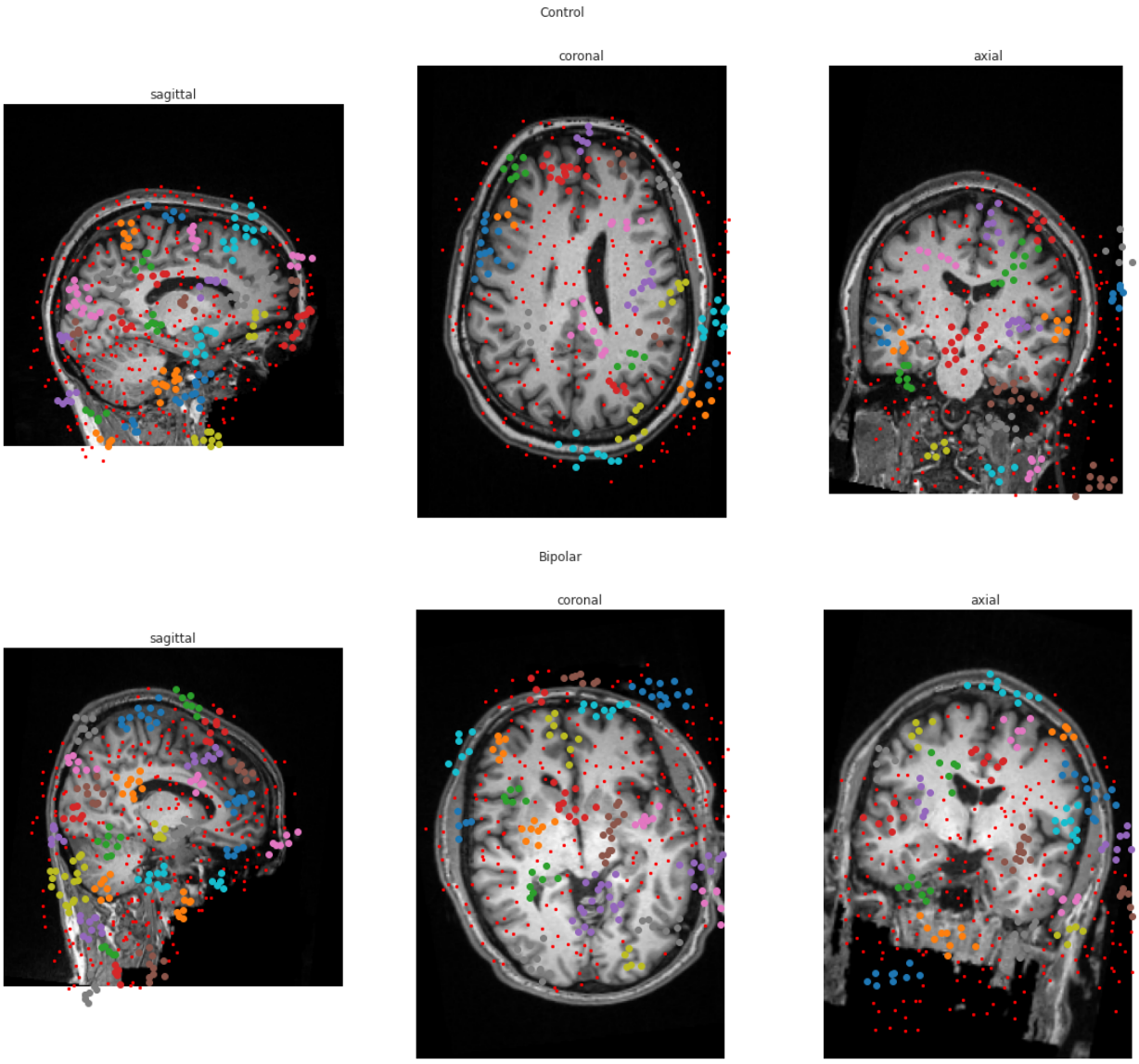

Figure 8. Clustering for Control and Bipolar from UCLA Dataset in model version 2. Clustering from dVoxResNet models trained on all four classes. Offset clusterings are shown on a random Control subject and a random Bipolar subject. The colors of different dot groups indicate groups of close offset points. The color is only to distinguish groups from one another.

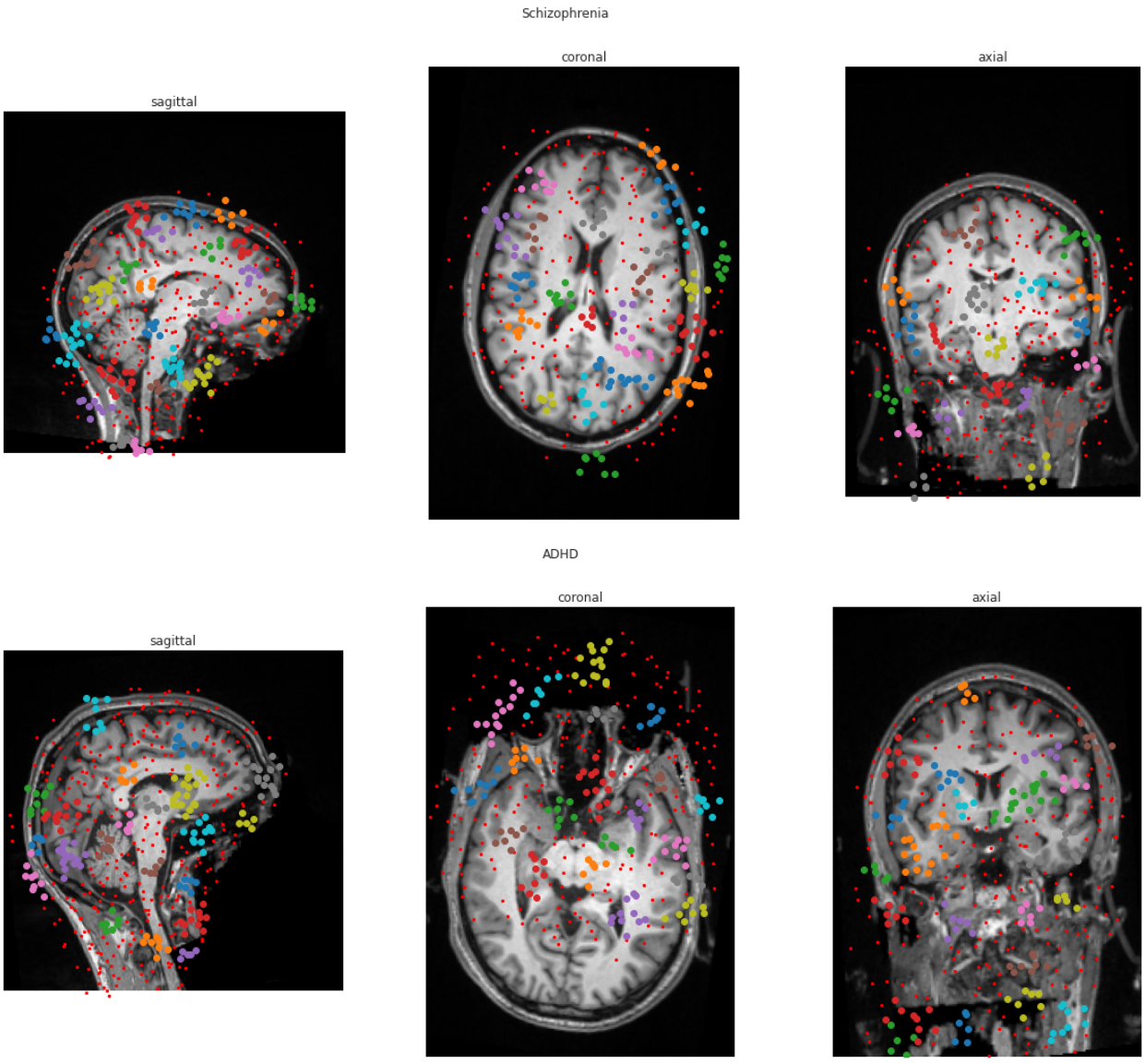

Figure 9. Clustering for Schizophrenia and ADHD from UCLA Dataset in model version 2. Clustering from dVoxResNet models trained on all four classes. Offset clusterings are shown on a random Schizophrenia subject and a random ADHD subject. The colors of different dot groups indicate groups of close offset points. The color is only to distinguish groups from one another.

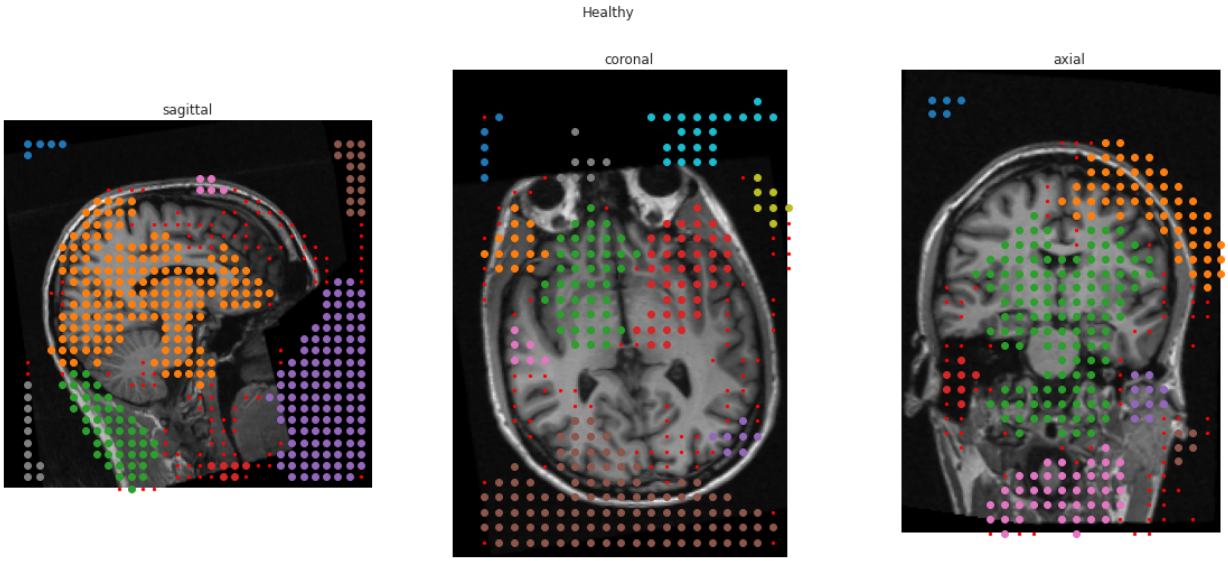

Figure 10. Clustering for a random healthy subject from NDA dataset.

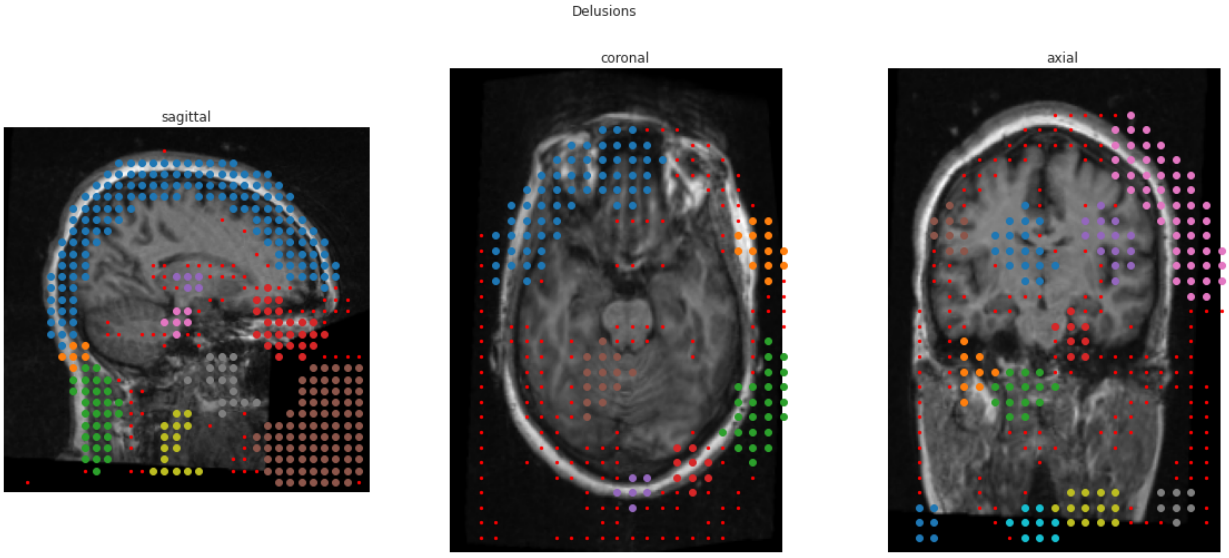

Figure 11. Clustering for a random subject with delusions from NDA dataset.

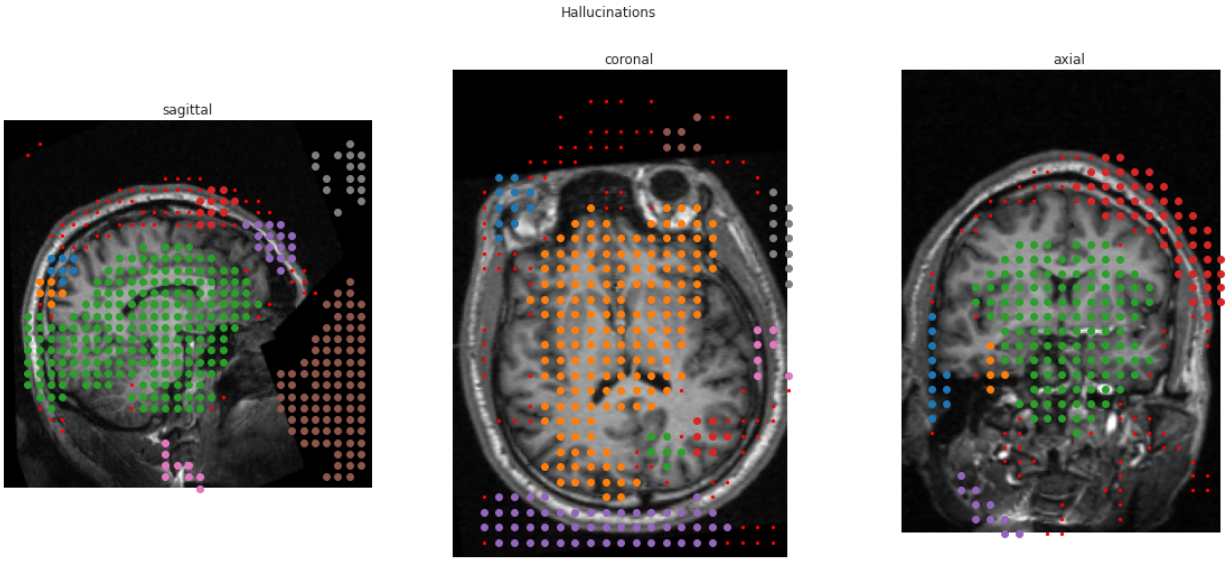

Figure 12. Clustering for a random subject with hallucinations from NDA dataset.

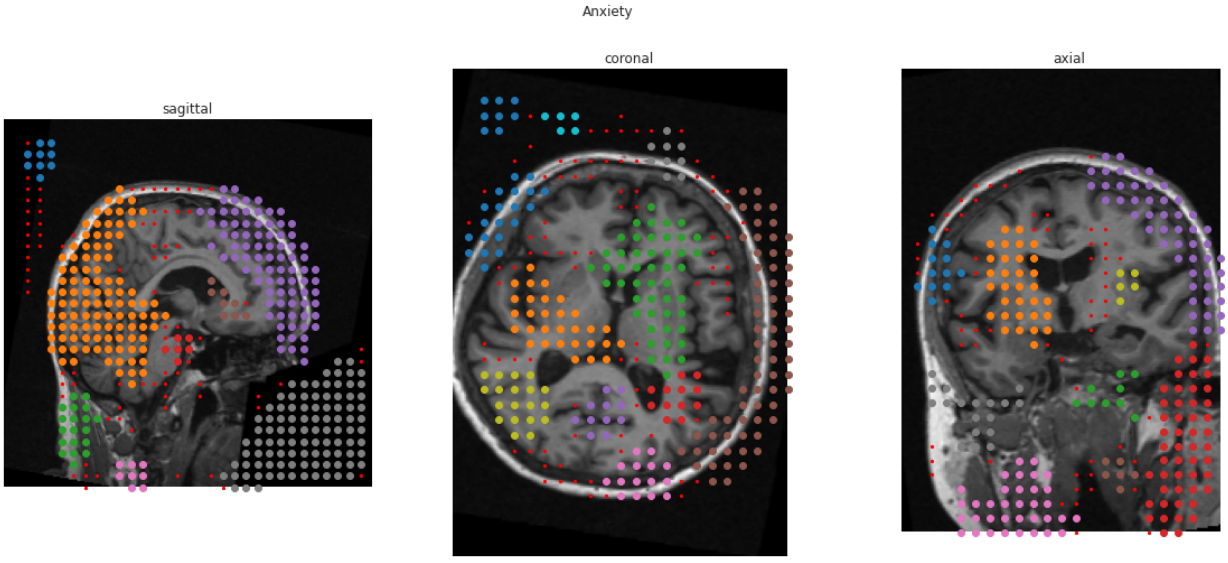

Figure 13. Clustering for a random subject with anxiety from NDA dataset.

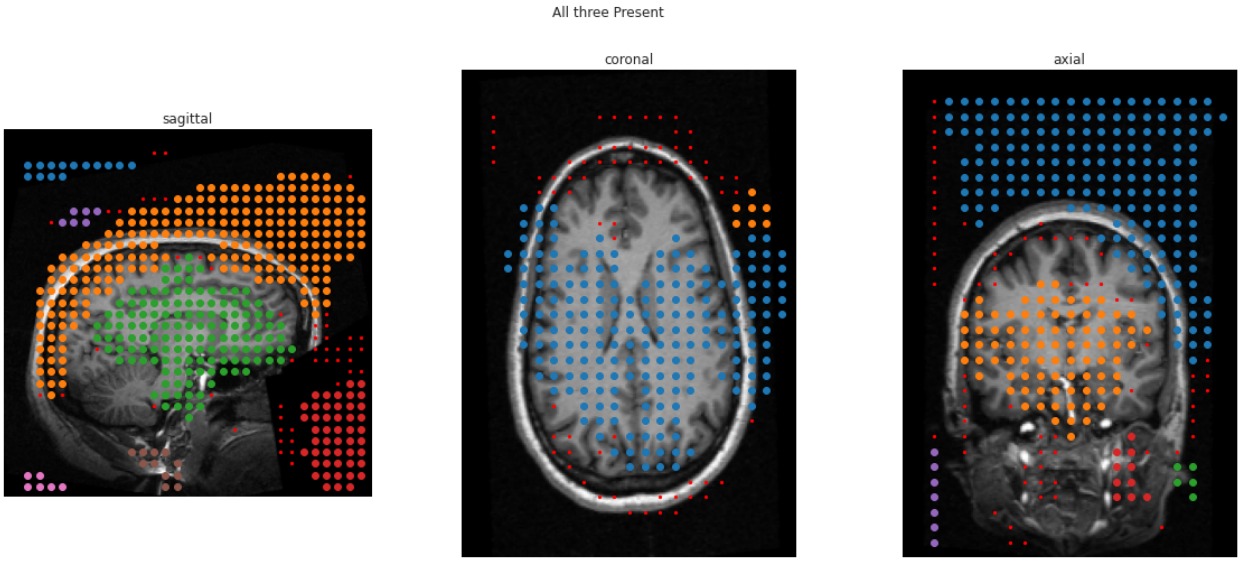

Figure 14. Clustering for a random subject with delusions, anxiety, and hallucinations from the NDA dataset.

Atlas segmentation as seen in Figure 15 was also done. Due to a limited color space of the images they were not overlaid on the existing dot structures to make them easier to read. However, they still provide unique information on the direct brain structure which dots may relate to.

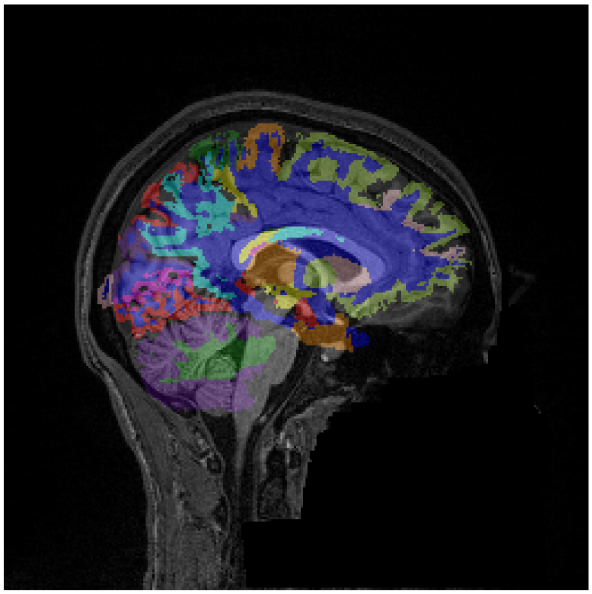

Figure 15. Atlas brain segmentation from a random sample in the UCLA dataset.

Grad-cam saliency attention maps for all layers of version one of the deformable vox resnet, can be seen in Figure 16. Clearly layers higher up in the model have the most usable views. They appear to highlight special brain regions. The lower layers by contrast learn less interpretable features. They highlight what could be the background or an unclear blob. It is likely that adding visuals via interpolated stacking of images may be useful. Through a multilayer view some of the lower layer features may have the context needed to make them useful.

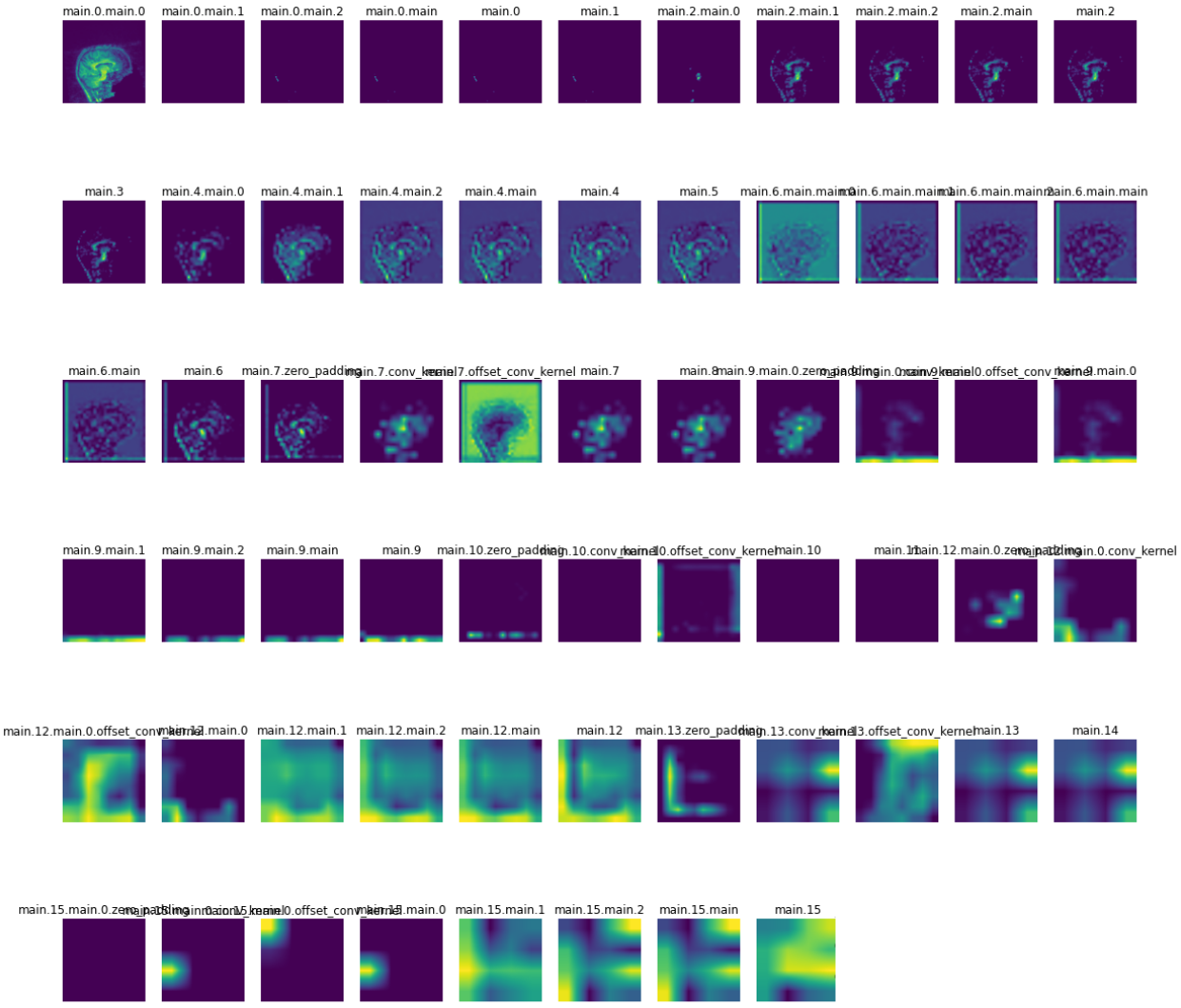

Figure 16. Grad cam visualizations from dVoxResNet.

3.4 Discussion

Offset displacement on both interpretations of the model show clear attention towards the brain. This demonstrates at a post-hoc level that the model finds specific parts of it more salient than others. While these results are qualitative at the moment it appears in both cases to pay more attention to the temporal lobe, auditory area, frontal lobe, and occipital lobe. These presences mesh with understandings of the resultant symptoms. Overall segmentation appears roughly homogeneous between the brain hemispheres. This indiacts at first pass in the symptoms and disorders examined the structural impact of them does not involve side specific or favoring functions.

Results on the NDA dataset show that the clustering of the offsets on the brain don't show as many meaningful patterns from qualitative analysis. While there are clusters on regions of the brain as shown in the figures above, the offset clusters are scattered across the brain with limited patterns that we could determine from qualitative assessment. In the future, we hope to aggregate all the offsets from each of the samples and do a statistical analysis to determine clusters that occur for each symptom over all the samples. It is possible that due to multi-label classification some of the more sparse sample types are not able to propagate offset displacements as much within the training time.

Results from the UCLA dataset in terms of clustering were much more promising. Clear areas of the brain were favored in a sparse way. Actual quantitative evaluation of these areas to tie them to an individual disorder are pending. Likely the improved performance was due to the binary training the model received in this case which would allow more information to reach higher into the model faster in training.

Given these results it is possible to clearly gain limited significance of brain attention areas in structure. Explainability results clearly offer a deeper insight into the model and derive greater inspection of how it is operating which can inform future model design. Given our results at this moment an ensemble based classifier where the comparies form healthy to each label / disorder appears to be both the most promising from an accuracy level and from an interpretability one. Further quantitative analysis of these model interpretations are needed to make stronger claims on the matter.

4 Bioethical Considerations for Medical AI

4.1 Introduction

With the rapid improvement of AI techniques, the incorporation of medical AI into the clinical setting is becoming closer to reality. Consequently, ethical concerns surrounding AI models need to be considered and addressed to facilitate implementation and ensure that the powerful benefits can be realized while also minimizing risks.

The medical field is guided by four fundamental principles: beneficence, non-maleficence, autonomy, and justice. These principles should also apply when developing AI models intended for the clinical setting. Beneficence provides the obligation for the clinician to act for the benefit of the patient (Floridi et al., 2018; Varkey, 2020). This principle may be satisfied in intention, where developers and clinicians aim to develop and use AI to improve patient care, but the beneficial impacts on patients may be overshadowed if the other ethical concerns are not addressed. Non-maleficence provides the obligation for no harm to be done to the patient (Varkey, 2020). For AI models, this relates heavily to the principle of justice, which calls for fair, equitable, and appropriate care, due to the possibility for biases to be introduced throughout development and deployment. Medical AI has the potential to reduce disparities in access and quality of care, so addressing biases is therefore necessary both to achieve equitable care and to reduce the potential consequence of patient harm. Finally, autonomy states that patients should be able to exercise self-determination.

Methods for ethical development of AI models have been widely investigated (Vokinger et al., 2021). In addition to efforts by developers, ethical usage of AI may also be promoted through regulation to establish baseline expectations. Like other medical equipment, implementation of AI/ML systems as supplemental medical tools will require approval and regulation by a centralized body. Within the US, this falls within the jurisdiction of the Center for Devices and Radiological Health (CDRH) branch of the US Food and Drug Administration (FDA), and more specifically, with the Digital Health Center of Excellence. At present, over 300 AI/ML-based devices have been approved. However, there are no specific regulatory pathways for AI/ML systems, so these devices have instead been cleared through one of three generic device pathways: premarket approval, de-novo premarket, or 510(k) (Muehlematter et al., 2021). The FDA has begun to lay the groundwork for specific regulation with its fairly recent creation of Software as a Medical Device (SaMD) categorization (Muehlematter et al., 2021). SaMD is defined as software which is intended for one or more medical purposes without being a part of a hardware medical device. In April 2019, the FDA released a discussion paper acknowledging that AI/ML-based SaMD is distinct from other products they oversee and that new regulatory policies must be developed to promote safety and effectiveness of these products (US Food and Drug Administration, 2019)).

The potential for variability of results due to biases is unique to these systems. Within the context of ML systems, bias is defined as systematic errors in outcomes for select subgroups of

the overall population (Vokinger et al., 2021). These biases can arise at multiple points throughout the model development and deployment process. This can lead to further concerns due to the ability of AI/ML models to iteratively improve by adapting to inputs from real-world use. While the dynamic nature of AI/ML-based SaMD can be beneficial, it is important for there to be regulation regarding these changes to ensure they limit biases.

As such, additional steps should be taken before, during, and after development to ensure trustworthy results and minimize the potential consequences of biased results. To accommodate the unique nature of ML-based SaMD, the FDA has proposed the total product life cycle (TPLC) regulatory framework to monitor these steps, while enabling rapid and regulated changes of the model post-release for optimization of patient health (US Food and Drug Administration, 2019). The following recommendations complement the TPLC framework and are meant to provide considerations for ways in which developers and regulatory bodies can reduce harm from AI/ML models: (1) Data disaggregation to produce representative datasets. (2) Data diversity reporting to increase and standardize transparency model characteristics to clinicians for informed decision making (3) Regulation of models to ensure applicability of models to contexts with differing resources. (4) Evaluating patient-feedback to continually monitor all aspects of informed consent and to execute continual regulation post-implementation.

4.2 Contextual Bias

The accuracy and reliability of ML systems depends heavily upon the data used for training during development. As a model is clinically implemented after testing, differences in datasets and available resources between development and deployment can result in different outcomes and impose potential clinical risks, also known as contextual bias (Price, 2019); (Finlayson et al., 2021). Non-representative data can lead to inaccurate outcomes such as a failure to properly diagnose a disease, leading to a lack of necessary treatment (Vokinger et al., 2021). For example, a model trained on predominantly white patients may underperform on underrepresented racial or ethnic groups (Finlayson et al., 2021). Alternatively, demographic shifts in the data could be possible, such as in the case of a hospital merger resulting in the addition of rural patients to an urban patient population, which may result in lower accuracy of model results (Finlayson et al., 2021). Similarly, a change in equipment used to collect data for the model can impart inaccurate results as the accuracy of the model may only extend to a specific version of equipment.

4.2.1 Creation of representative datasets

Many developers are able to access data on open sources such as SchizConnect and OpenNeuro. Here, they are granted the ability to specifically sort for data, and apply labels that filter for specific populations. However, how useful is this filtration when much of the data is already inclined towards a population? In other words, is the data truly diverse and reflective of the general population if the data as a whole contains demographic information from only a

select group of people? This leads back to the issue of contextual bias, but more specifically selection bias. Selection bias occurs when selected patient samples are improperly randomized or not at all and are thus, unreflective of the general population. Biases such as these carry downstream negative effects onto minority populations when research is translated into practical application. Developers are realizing that training and testing datasets are not the most demographically representative of target populations and are non-representative (Manrai et al., 2016). For example, databases with skin lesions and melanomas used to train certain medical AI lack images from patients with darker skin tones (Adamson & Smith, 2018). Thus, to combat contextual biases like selection bias derived from demographic disparities, disaggregating data to create more representative datasets targets the data collector rather than the developer. This portion of the paper suggests a solution to data bias while considering the technical and ethical facets of AI specifically when used in an imaging setting.

We propose creating more representative datasets through data disaggregation to analyze the diversity of compiled datasets in order to inform data collection efforts if certain demographics are found to be underrepresented (Vokinger et al., 2021). Racial and ethnic disparities and the resulting inequities can be minimized if there is a diverse, high-quality pool of data that can guide analysis. "Data disaggregation", for the purpose of this investigation, will thus be referred to as the purposeful division of collected data into its demographic subcategories. Many times, epidemiological studies are divided into a few separate patient outcomes: "White", "Black", or "other" (Cahan et al., 2019). However, the reality is that the vast majority of the nation is more diverse. For example, when referring to the Asian American community, what lies under that are more than 60 different ethnicities and over 100 languages by definition. From the developer's point of view, data disaggregation enhances the precision and outcome of the algorithm.

Justice, as a clinical ethics principle, must be upheld in the clinical setting despite AI integration. In other words, differential diagnoses informed by an AI must achieve fair and equitable treatment of the respective patient. To achieve this, the AI must objectively be taught equity in the form of equal representations of subgroup data. During surveys or data collection, the data collector/investigator must conjure a well-rounded demographic survey that is a true representative population. For example, the National Health and Nutrition Examination Survey conducts selective surveys that involve interactive interviews, and physical examinations for carefully selected sample populations that are representative of the general population. They select a highly specific population via the census and incentivized national surveys. More importantly, the populations they target are highly diverse and ethnically distinct. For reliable, realistic statistics, NHANES over-samples individuals who are older than 65 years old and minority populations like "African Americans, Asians, and Hispanics". They essentialize their data to represent a highly diverse set of populations reflective of the actual average demographics of the nation in order to accurately reflect the health status indicator estimates for particular subgroups (Zipf et al., 2013). NHANES therefore makes the case for sub-data

disaggregation in their process of specifying highly disaggregated populations in a holistic demographic classification.

Machine learning uses regulated data as its foundation for elucidating patterns and eluting predictions. Therefore, sub-data disaggregation provides a means to improve health research and policy making. The exact recommendation has the data collector sequentially divide categorical demographic data into sub-demographic data within race, ethnicity, and age corresponding to a true, diverse reflection of the community in mind. The proposed process to disaggregate data is stratified in a sequential fashion. Geographical survey and sample locations should be chosen with the intention to increase data collection of higher risk and minority populations (*Obesity and Hispanic Americans*, 2021). In other words, due to typical oversight of at-risk and minority populations, special attention should be given to these communities to account for ever-growing national diversity. For a more informed sample, participants should be either provided with a physical check-up or have an existing, updated medical history. With the chaotic disorganization seen in EHR systems today, up-to-date information is vital to eliminating implicit biases. Prior to the aforementioned steps, prevalence of the disease/illness/condition should be epidemiologically or sociologically tracked per state per county. This will heavily inform the location of sampling. Sampling regions would be selected with higher proportions of individuals within certain sub-demographic groups. In these regions, patient samples are recruited and selected at random, with detailed consent. Here, when the selected region has a larger proportion of a certain minority, sampling for those populations rises to much higher probability. Persons would be chosen at random from each household while being mindful of gender/sex/ethnicity divisions. This methodology would collectively allow for investigators to select demographically-balanced samples. Ultimately, these simple, but specific criteria work to ensure that data is properly distributed, representative, and builds credibility to the equity and justice of the respective study.

4.2.2 Increased transparency through the report of training demographics

To endow clinicians with the ability to make fully informed decisions regarding appropriate clinical usage of AI/ML models, transparency of data used for the development of the model should be ensured. With a limited understanding of the basis for model outputs, it is difficult for subsequent clinical decisions to be informed and to mitigate potential risks. Despite efforts to enhance the explainability of models (refer to Section 3.1 for a description on "explainability"), there is an immediate need to provide clinicians with further information regarding the model to guide the capacity in which the model is used for patient care.

Currently, clinicians are limited to various studies which apply machine learning methods to clinical prediction and diagnostic modeling. However, the unique risk factors of AI/ML models within these studies are not effectively reported. Recent efforts to coordinate the reporting protocols of clinical prediction models using AI have led to the development of Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis (TRIPOD) statement and the Prediction model Risk Of Bias Assessment Tool (PROBAST) (Wolff et al., 2019). Reporting criteria that assess the multi-dimensional biases that come with

ML can help regulatory bodies provide better guidance. Stakeholders and end-users may better understand whether ML is suitable for their purposes and may dictate usage in a clinical setting (i.e. a patient with a highly-complex diagnosis may benefit with ML). While tools like PROBAST enable assessment of the risk of bias and model empiricism, they do not enforce reporting of their respective conclusions on model characteristics such as participant diversity, outcome definition, predictors, etc. (Jong et al., 2021).

Reporting is a gateway into transparency. In their discussion of AI/ML-based SaMD, the FDA recognizes transparency and real-world performance monitoring as necessary components of their total product lifecycle (TPLC) regulatory approach (US Food and Drug Administration, 2019). Currently, details provided by developers on the nature of data usage, algorithm development, and products vary wildly (Richardson, 2021). For example, an analysis revealed that only 1 of 10 FDA-approved AIs for breast imaging provided demographic information (i.e racial, etc.) as a justification for product validation (Ross, 2021). Having disclosure on the demographic information and additional aspects of training data can guide providers to select a product most fitting for their clinical environment keeping contextual bias in mind. This call for transparency in the disclosure of data used by AI/ML-based SaMD has been echoed by the American College of Radiology (ACR) and the American College of Surgeons (ACS) to help clinical decision making and increase trust in AI/ML in clinical use (Schneider et al., 2021). As SaMDs grow, patient and clinician communication are becoming increasingly important. Patient understanding is especially key in informed consent - for the patient to know the whole truth behind their treatment and care including AI usage (refer to *Section 4.3: User feedback*). Facilitating clinician understanding of the model is essential to support informed decision-making for clinicians and patients alike.

Providing a comprehensible report of demographic and contextual characteristics of the data used in the model to clinicians will enable informed decision-making for patient treatment. Ideally, regular reports of data diversity should be reported to accommodate evolving SaMDs and the model lifecycle (Schneider et al., 2021). To aid understanding of the model testing data, the report should outline characteristics of the population demographics including age, sex, race, ethnicity, gender, and geographic region. Having this knowledge can inform users of the potential limitations a model may have with different datasets. Clinical characteristics of the data should be included such as the prevalence of comorbidities, treatment plans/status, and timing/onset of data collection. Furthermore, the population demographics could be disaggregated on the basis of clinical characteristics to highlight aspects of the data which are not palpable. For example, if two ethnicities are equally represented within a dataset but differ in clinical characteristics, such as prognosis, this finding should be included in the report. It is also necessary for the report to include domain characteristics of the data including the number of facilities included in the set and the manufacturer and model of data acquisition devices. Efforts at Duke University and the Mayo Clinic have worked to create a label (similar to a nutrition label) that would describe how the model was developed, tested, and deployed (Brodwin, 2020; Jercich, 2022; Sendak et al., 2020). Motivated by this, we propose that reporting of dataset

information be integrated into model labeling alongside other model aspects. The diversity label should clearly demonstrate within what level of context the ML model operates. The information would be easily integrated into existing EHR systems, such as CRISP, and would be easily accessible.

4.2.3 Regulation of usage environments

The FDA should also seek to regulate the contexts in which models may be implemented. Differences in available resources in development environments introduce increased risks for patient harm. Resources may include the amount of healthcare personnel and their levels of expertise, amount and sophistication of available equipment, and available drugs.

When implementing a model in a specific location, consideration of the resources and capabilities of the environment are necessary to minimize risks of inaccurate outcomes. Models developed for treatment recommendation are of particular concern. If model training and evaluation occur in high-resource environments, the model is likely to recommend treatments appropriate for similar environments. However, such treatments may be inappropriate, unfeasible, or of higher risk than other treatments when considered within the context of a lower-resource environment (Price, 2019). As an example, consider the treatment of malaria. Though malaria is unlikely to require analysis from an AI model, it provides a direct example of the effects of contextual environments on applicable treatments. Treatment of malaria in most parts of the world typically involves nonsteroidal anti-inflammatory drugs (NSAIDs) to address malarial fever (Weissglass, 2021). But within the low-resource African regions, where malaria occurs with much higher incidence, NSAIDs are not recommended as they may result in stomach ulcers and kidney damage if taken without food or water (Weissglass, 2021). Other possibilities for less-than-ideal treatment outcomes may be a procedure requiring unobtainable equipment or personnel with more expertise than is available, and drugs that are not available due to cost or storage requirements.

Restriction of model usage to specific contexts can reduce risks associated with contextual bias, but it will likely have the additional consequence of preventing low-resource environments from ever having access to the power of medical AI models. AI models present a powerful tool for the democratization of healthcare by allowing for the knowledge of an expert to be leveraged in contexts where the maximum level of expertise and the size of the healthcare workforce are diminished (Weissglass, 2021). Thus, while risks of consequences from bias present conflicts with non-maleficence, inequitable access presents conflicts with beneficence and justice.

Regulation should balance the need to minimize risks with the need to allow the benefits of models to come through. As such, intended usage contexts should be restricted, but regulation should encourage developers to expand initial contexts to include low-resource environments. According to the FDA's TPLC framework, developers will have the option to submit modifications to original intended usage contexts (US Food and Drug Administration, 2019). Though full generalizability to all contexts is ideal, proving extensive generalizability will be a

difficult and time-costly process due in part to the likelihood of insufficient data from low-resource environments (Price, 2019). Rather than initially requiring full generalizability or leaving applicability to multiple contexts as an option, regulation could require that developers submit plans for implementing such modifications after initial review. This approach would therefore not completely prevent models from benefitting any contexts while also supporting the safe expansion of usage contexts.

4.3 User feedback and Informed Consent

User feedback in the context of human-centered design is critical for maximizing the benefit of the user-experience and improving the underlying output of the model. Feedback allows the creators to have a real sense of the impact that their model has on the users themselves, and also provides the model creators with qualitative and quantitative metrics that they can use to develop their personal goals. In the context of healthcare-associated AI devices, the use of patient-feedback to continually evaluate through the TPLC, especially post-implementation, can be just as important. Indeed, patients are the central beneficiaries of AI-based clinical tools and medical applications, and although such tools may help improve diagnostics, treatment planning, and patient outcomes, patients should be ensured that they will not be harmed in any way by AI-based devices, and rather will benefit for the use of the technology in the healthcare setting. To enable the integration of the technology into healthcare, along with device developers seeking CMS approval, the associated concerns and risks should be addressed through patient-surveys and discussion of possible concerns with the patient. To measure such feedback and evaluate informed consent, the following is recommended: (1) surveys/interviews with patients using the specified health services, (2) resourceful surveys with clinicians, and (3) close observation of clinical encounters between patients and clinicians. Patient comments at the end of surveys can also help enhance patient experience and healthcare delivery, and such cautious actions can serve to help measure patient-centered care (Silva, 2014).

4.3.1 Patient Concerns

Concerns with the technology itself, with ethics, and the regulation are associated with the perceived risks of the technology. For the technology, the perceived concerns include a lack of transparency, complexities associated with interpreting results, and the safety of recommendations driven by the AI. There is performance anxiety in that the users may have a perception of the threat for the system to malfunction and not work as intended, rendering it unable to deliver the service (Esmaeilzadeh, 2020). Systems can be vulnerable to hackers (untraceable attacks) and unexpected errors. These medical errors could endanger patient safety, and so, users might be concerned that the AI mechanisms can lead to inaccurate diagnoses/treatments. Nonrepresentative datasets and incompleteness in the models can also produce medical errors or inaccurate predictions. There are also perceived barriers to communication as the use of AI devices in healthcare may also change communications between

patients and physicians. Ultimately, privacy concerns, mistrust in the AI mechanisms, and social biases may also influence how a patient feels about such technologies. Patients are often also concerned about cost, and whether they can gain access to such technologies via their insurance plans. There are also concerns with the regulations and governing of the AI systems themselves. There are perceived unregulated standards, and this is a critical challenge to the AI tools not yet being transparent. As the technology is rapidly developing, helping patients and clinicians understand how the device works, and what it is used for is a critical part of ensuring transparency. Furthermore, algorithms that continuously change with features may not be able to abide by original clinical trials. The current standards should be formalized to maintain the safety and impact of AI-based healthcare technologies. Regulatory agencies must agree on a set of standards that establishes official requirements, policy, and guidelines, such as that offered by the FDA. Perceived liability is also an issue as there are concerns about who will account for the errors that may occur with AI-based decisions. Autonomous decisions by AI-based healthcare devices creates a risky situation where it isn't clear who is responsible for wrong recommendations. Also, there is a liability risk if recommended treatment options by AI are dismissed. For patients to be able to use and trust such technologies in a healthcare setting, these perceived risks must be addressed. User-feedback and ensuring informed consent by facilitating communication among the patient-clinician, developer-clinician, and developer-patient relationships is critical for identifying actionable plans against such perceived risks.

4.3.2 FDA Action Plan

The Digital Health Center of Excellence falls within CDRH of the FDA, and it aims to align interests for digital health innovation as well as provide support to all stakeholders involved including the developer community. In October of 2021, the Digital Health Center of Excellence held a meeting with the Patient Engagement Advisory Committee (PEAC). The PEAC consists of patient advocates, caregivers, and patients, and this committee along with the FDA helped outline a five-point action plan for AI/ML-enabled medical devices (Kiarashi et al., 2021). As mentioned, there are currently more than 300 FDA-authorized AI/ML-enabled devices, and so the action plan consisted of: (1) updating a tailored regulatory framework for AI/ML-based SaMD, (2) encouraging GMLP through development, (3) supporting transparency through a patient-centered approach and enhancing trust, (4) supporting methods related to model evaluation, improvement, and robustness as well as identifying bias and eliminating it, and (5) working with stakeholders through the real-world performance. Increasing transparency and trust remains an important part of addressing potential patient concerns.

4.3.3 Patient Surveys

Chats, surveys, emails, and support tickets are just a few examples for how developers might traditionally provide feedback, providing a means of communication for the developer-patient as well as patient-clinician relationships. With regards to healthcare delivery, patient feedback can be in the form of surveys that incorporate questions about the firsthand or

secondhand experience of the patients with the model. The surveys may ask, for example, whether the physicians explained the purpose of the diagnosis/treatment or whether the medical staff is responsive when patients need them. The primary goal of this survey would be to have comparable data from the patient's perspective on the care that they receive. A few commonly used tools for measuring patient-care include the Individualized Care Scale (ICS), the Measure of Process Care (MOPC), the Person-centered Care Assessment Tool (P-CAT), and the Person-centered Climate Questionnaire (PCCQ). The ICS serves to evaluate how patients perceive individuality in the care and their evaluation of the overall healthcare experience. The ICS and such perception surveys can be particularly important tools for data developers to consider because it provides an opportunity to be an integral part of the two-way dialogue between the patient and those contributing to the treatment of the patient (Kiarashi et al., 2021). Whereas customer surveys have an emphasis on the satisfaction for the products/services they engage with, patient surveys have an emphasis on the patients' perceptions of the care that they receive, often through check-box questions. The ICS in particular is a 15-minute questionnaire with 40-items that is a promising tool that provides an opportunity for the expression of patient perceptions of care (Suhonen et al., 2005).

4.3.4 Informed Consent

The level of trust that patients and clinicians have also directly has an impact on how much the users will be able to rely on the model. This means that this further influences the efficacy of the health care diagnosis. If a particular development is accepted into the clinical community, the patient might have to accept it de facto based on the trust that the patient has with his/her provider. Also, it is important to note that this level of trust may not always have a positive correlation with the patient outcomes. Thus, it is quite important that the user experience is improved and addressed with any healthcare associated AI device so that patients and clinicians can feel that they can trust or use such technologies. Ensuring that there is communication between the developer and clinician is also important to equip clinicians with the necessary tools for understanding the recommendations made by the AI-healthcare devices. As the PEAC committee mentions in their discussion, device developers must properly find ways to communicate how the device works to the clinician, and must also help the clinician find ways to explain the processes to the patient (Kiarashi et al., 2021). Knowing how it works as well as the extent to which it works is important to identify when the associated device might not be working how it's supposed to. Understanding how it works is also important for identifying bias, and so finding ways to communicate as well as educate should be considered to help improve that transparency.

4.3.5 CMS Approval

Furthermore, something that shouldn't be overlooked comprises the economics/costs of such technologies. Whether a clinician will use the technology might be dependent on its performance but also on whether that type of care is reimbursed. Doctors will not reimburse for

the care and will not use the AI technology if there is no reimbursement. After FDA approval, the device makers should also obtain reimbursement approval from the Center of Medicare and Medicaid Services (CMS) for use of the products by Medicare beneficiaries. The approval for CMS to provide reimbursement for use of the products with the indication can be challenging yet could be a critical step for acceptance into the community. The cost of using such devices in the clinic and how the patients will be billed is crucial to consider because it may determine whether the devices are adopted for clinical use in the first place.

There are overall several considerations to take into account when emphasizing ways to help enhance patient trust and facilitate communication among patients, clinicians, and developers. Emphasizing post-implementation within the TPLC and finding ways to continue to have model feedback going from the patients to the clinicians to the developers should help increase that trust/transparency. In addition, when considering this patient-centered approach, communicating about data privacy, how the AI/ML-enabled device works, and the device's intended use will be an important aspect of patients being able to make informed decisions about their care. Finally, techniques to reduce the burden of economics and costs of the device through such CMS programs will help integration of the technology into the healthcare system.

5 Conclusion

Efforts to better understand mental illnesses have been ongoing. The National Institute of Mental Health (NIMH) began its Research Domain Criteria (RDoC) project with the goal of understanding mental disorders based on observable behavior and neurobiological measures (Cuthbert & Insel, 2013). This is accomplished through the investigation of multiple units of analysis (ie. genes, molecules, cells, circuits, physiology, behavior, self-reports, and paradigms) (Insel & Cuthbert, 2009).

Our work for the past three years focused on learning more about mental illnesses, developing algorithms to detect physiological changes in the brain, and addressing the ethics behind real world use of these algorithms. While there are many studies linking changes in the brain to mental illnesses, we decided to improve upon these methods and develop visualizations to make them interpretable (Wilczyńska & Waszkiewicz, 2020). By using deep learning, we would be able to detect changes in sMRI between patients with mental illnesses and those who were healthy. In order to begin our research, we investigated prior applications of deep learning in mental illness diagnosis, including Alzeimers disease, Major Depressive Disorder, Schizophrenia, Bipolar Disorder, and ADHD.

In our first implementation, we decided to use supervised and unsupervised models and compare the biomarker results that were produced from both approaches. Our goal was to develop an accurate model on the data using the supervised method, and an unbiased model of the biomarkers using the unsupervised method. While we did get some results for the supervised method, the unsupervised method was very challenging to train and we ran into many issues that ultimately led to us moving on to a different approach.

Our second implementation was to improve on our supervised model by developing on recently published work on using deep learning for MRI scans (Carreira & Zisserman, 2018; Yang et al., 2021). We also developed a novel network using Atlas Attention maps. These attention maps segment the brain into multiple areas which can help with visualizing which areas of the brain were more important when deciding the disease. We also increased our dataset size to include symptom based classification, by obtaining access to NIMH NDA datasets, so that symptoms, instead of diseases, would be predicted from brain scans. From these results we saw that atlas attention was able to improve or maintain performance across three classification tasks. We were also able to train a model on the NDA dataset which was able to detect the three symptoms with 50.84% accuracy.

While most deep learning methods for medical applications are currently still in development, continuing advancements mean that implementation into clinical settings are coming closer to reality. We identify various recommendations to address bioethical concerns relating to the effects of bias that should be taken into consideration when developing and deploying deep learning models. The recommendations are meant to inform, guide, and provide potential considerations for AI/ML SaMD devices to minimize the risks of contextual biases and increase transparency throughout the TPLC of a medical device. They will equip healthcare

providers with the tools to understand potential risks, provide ways to enhance patient-centered care, and ensure continual regulation/updating of the model post-implementation are all crucial aspects to help facilitate the relationships among all stakeholders: clinicians, patients, developers, and regulators.

5.1 Future Work

Our results emphasize the difficulty of distinguishing between multiple disorders and multiple symptoms. This task becomes especially difficult when limited to using only sMRI data. The potential of other imaging and imaging-like techniques, such as EEG, MEG, and fMRI, should continue to be investigated. Additionally, for certain psychiatric disorders, other non-imaging features can provide valuable diagnostic information, for instance, prior history of substance abuse. An ideal classifier might integrate imaging data with other non-imaging data modalities through techniques such as ensemble classification. Future studies that use similar pixel processing methods of deep learning would gain greatly from using a multi-image format such as T1 along with T2 weighted images for a patient as each image form would encode unique brain information that is directly interrelated but not visible in full scope without the use of this multi model approach.

Additionally, the avenue of model explainability can go much further within brain image classification. Collecting the sparsity of groupings of attention encodings, offsets, or other model decision metrics could help indicate biomarkers and should be studied at a class wide base to find repeating features. Adding metrics like KAR or RAOR in addition to multi method performance could also open more anquenuse to evaluate the model perforce across classes with than just f1 score or accuracy. If computation methods are going to enter medicine the results the models get and how they got there will need to be interpretable to the clinician. Thus metrics of its interpatiblity should be investigated for both research and use purposes.

Like many previous studies using neuroimaging data, we were limited by the sizes of our datasets. Deep learning models generally require large training datasets in order to generalize well. The amount of data required generally increases as the number of trainable parameters increases, making this consideration especially important for large imaging models such as those used to conduct our experiments. We hope that future work in this area will benefit from larger, cleaner datasets like those being compiled by organizations such as ADNI and OpenNeuro. Ideally, these datasets will make distinctions between medicated patients and drug-naive ones. Consistent with our ethical recommendations in Chapter 4, we hope that future work in this area will continue to push for representative datasets and transparency in data reporting in order to minimize bias. An alternate avenue of research would be more detailed data augmentation. Existing machine learning techniques like general adversarial networks could be used to take sparse samples and effectively rebalance data sets. This may also open the door to dense brain encodings that could allow image samples to be more tersely summarized and thus possible brain image features to be discovered via the reconstruction process.

Independent of classifier performance, many obstacles such as image modality, explainability, and data quantity, accessibility, and bias exist to the potential widespread clinical application of deep learning-based diagnostic tools. Research in these areas will serve as a continuation to the work done in this project and more importantly as candidates for future endeavors in this field. We hope that developments in the aforementioned areas will continue, and improve existing technologies in order to better combine the power of computer-aided diagnostic tools with human experts and clinicians.

6 References

[1705.02315] ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. (n.d.). Retrieved January 28, 2022, from https://arxiv.org/abs/1705.02315

A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder—PubMed. (n.d.). Retrieved January 28, 2022, from https://pubmed-ncbi-nlm-nih-gov.proxy-um.researchport.umd.edu/27779627/

Akkus, Z., Cai, J., Boonrod, A., Zeinoddini, A., Weston, A. D., Philbrick, K. A., & Erickson, B. J. (2019). A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence–Powered Ultrasound for Improving Clinical Workflow. *Journal of the American College of Radiology*, *16*(9, Part B), 1318–1328. https://doi.org/10.1016/j.jacr.2019.06.004

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's

disease. *Alzheimer's & Dementia*, *7*(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. *ArXiv:2002.00772 [Cs]*. http://arxiv.org/abs/2002.00772

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. *ArXiv:1806.08049 [Cs, Stat]*. http://arxiv.org/abs/1806.08049

Anderson, I., Haddad, P., & Scott, J. (2012). Bipolar disorder. *BMJ*, *345*. https://www.bmj.com/content/345/bmj.e8508.full

Ankerst, M., Breunig, M. M., Kriegel, H., & Sander, J. (1999). *OPTICS: Ordering Points To Identify the Clustering Structure*. 49–60.

Arbabshirani, M., Kiehl, K., Pearlson, G., & Calhoun, V. (2013). Classification of schizophrenia patients based on resting-state functional network connectivity. *Frontiers in Neuroscience*, *7*. https://www.frontiersin.org/article/10.3389/fnins.2013.00133

Artificial intelligence approach to classify unipolar and bipolar depressive disorders | SpringerLink. (n.d.). Retrieved January 28, 2022, from https://link-springer-com.proxy-um.researchport.umd.edu/article/10.1007/s00521-015-1959-z

Attention-Deficit/Hyperactive Disorder (ADHD). (2014). National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/attention-deficit-hyperactivity-disorder-adhd

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020).

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barton, N. T. L., Paul Resnick, and Genie. (2019, May 22). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Brookings*. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., & Filippi, M. (2018). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage : Clinical*, *21*, 101645. https://doi.org/10.1016/j.nicl.2018.101645

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Birkhäuer, J., Gaab, J., Kossowsky, J., Hasler, S., Krummenacher, P., Werner, C., & Gerger, H. (2017). Trust in the health care professional and health outcome: A meta-analysis. *PLoS ONE*, *12*(2), e0170988. https://doi.org/10.1371/journal.pone.0170988

Breijyeh, Z., & Karaman, R. (2020). Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *Molecules*, *25*(24), 5789. https://doi.org/10.3390/molecules25245789

CADDementia—Grand Challenge. (n.d.). Grand-Challenge.Org. Retrieved March 27, 2022, from https://caddementia.grand-challenge.org/

Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S., & Rubin, D. L. (2019). Putting the data before the algorithm in big data addressing personalized healthcare. *Npj Digital Medicine*, *2*(1), 1–6. https://doi.org/10.1038/s41746-019-0157-2

Campese, S., Lauriola, I., Scarpazza, C., Sartori, G., & Aiolli, F. (2019). Psychiatric Disorders Classification with 3D Convolutional Neural Networks. *Proceedings of the International Neural Networks Society*. https://link.springer.com/chapter/10.1007/978-3-030-16841-4_6

Carreira, J., & Zisserman, A. (2018). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *ArXiv:1705.07750 [Cs]*. http://arxiv.org/abs/1705.07750

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), 1–12. https://doi.org/10.1038/s41398-019-0607-2

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *The New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. *26th International Conference on Intelligent User Interfaces*, 307–317. https://doi.org/10.1145/3397481.3450644

Cobots in knowledge work: Human – AI collaboration in managerial professions—ScienceDirect. (n.d.). Retrieved March 26, 2022, from https://www.sciencedirect.com/science/article/pii/S014829632030792X

Cui, Z., Gao, Z., Leng, J., Zhang, T., Quan, P., & Wei, Z. (2019). *Alzheimer's Disease Diagnosis Using Enhanced Inception Network Based on Brain Magnetic Resonance Image*. https://doi.org/10.1109/BIBM47256.2019.8983046

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable Convolutional Networks. *ArXiv:1703.06211 [Cs]*. http://arxiv.org/abs/1703.06211

de Bruijne, M. (2016a). Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, *33*, 94–97. https://doi.org/10.1016/j.media.2016.06.032

de Bruijne, M. (2016b). Machine learning approaches in medical image analysis: From detection to diagnosis. *Medical Image Analysis*, *33*, 94–97. https://doi.org/10.1016/j.media.2016.06.032

Dolz, J., Desrosiers, C., & Ben Ayed, I. (2018). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, *170*, 456–470. https://doi.org/10.1016/j.neuroimage.2017.04.039

Donnelly, J., Barnett, A. J., & Chen, C. (2021). Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. *ArXiv:2111.15000 [Cs]*. http://arxiv.org/abs/2111.15000

Drukker, L., Noble, J. A., & Papageorghiou, A. T. (2020). Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. *Ultrasound in Obstetrics & Gynecology*, *56*(4), 498–505. https://doi.org/10.1002/uog.22122

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Evans, W., Morrill, M., & Parente, S. (2010). Measuring inappropriate medical diagnosis and treatment in survey data: The case of ADHD among school-age children. *Journal of Health Economics*, *29*(5). https://www.sciencedirect.com/science/article/abs/pii/S0167629610000962?casa_token=yeVnNR8Ww7oAAAAA:1TfZXZmMVRBE0l9MwO8L9b0FU1uZrgx-AhrxiOAfZSuLS8oE4q8BUv7zVguMC-KWFsj9XDCxQq8V

Explainability Methods for Graph Convolutional Neural Networks | IEEE Conference Publication | IEEE Xplore. (n.d.). Retrieved March 26, 2022, from https://ieeexplore.ieee.org/document/8954227

Falkai, P., Schmitt, A., & Andreasen, N. (2018). Forty years of structural brain imaging in mental disorders: Is it clinically useful or not? *Dialogues in Clinical Neuroscience*, *20*(3), 179–186.

Folego, G., Weiler, M., Casseb, R. F., Pires, R., & Rocha, A. (2020). Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI. *Frontiers in Bioengineering and Biotechnology*, *8*. https://www.frontiersin.org/article/10.3389/fbioe.2020.534592

Ford-Jones, P. C. (2015). Misdiagnosis of attention deficit hyperactivity disorder: 'Normal behaviour' and relative maturity. *Pediatrics and Child Health*, *20*(4). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4443828/

Fung, G., Deng, Y., Zhao, Q., Li, Z., Miao, Q., Li, K., Zeng, Y., Jin, Z., Ma, Y., Yu, X., Wang, Z., & Shum, D. (2015). Distinguishing bipolar and major depressive disorders by brain structural morphometry: A pilot study. *BMC Psychiatry*, *15*. https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-015-0685-5

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine Learning in Major Depression: From Classification to Treatment Outcome Prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037. https://doi.org/10.1111/cns.13048

Georgiou, T., Liu, Y., Chen, W., & Lew, M. (2020). A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision. *International Journal of Multimedia Information Retrieval*, *9*(3), 135–170. https://doi.org/10.1007/s13735-019-00183-w

Gotkowski, K., Gonzalez, C., Bucher, A., & Mukhopadhyay, A. (2020). M3d-CAM: A PyTorch library to generate 3D data attention maps for medical deep learning. *ArXiv:2007.00453 [Cs]*. http://arxiv.org/abs/2007.00453

Grotegerd, D., Suslow, T., Bauer, J., Ohrmann, P., Arolt, V., Arolt, V., Stuhrmann, A., Heindel, W., Kugel, H., & Dannlowski, U. (2013). Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: A pilot study. *European Archives of Psychiatry and Clinical Neuroscience*, *263*.

Guo, Y., Ji, J., Lu, X., Huo, H., Fang, T., & Li, D. (2019). Global-Local Attention Network for Aerial Scene Classification. *IEEE Access*, *7*, 67200–67212. https://doi.org/10.1109/ACCESS.2019.2918732

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, *69*, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

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

Health, C. for D. and R. (2020). What are examples of Software as a Medical Device? *FDA*. https://www.fda.gov/medical-devices/software-medical-device-samd/what-are-examples-software-medical-device

Henschel, L., Kügler, D., & Reuter, M. (2022). FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage*, *251*, 118933. https://doi.org/10.1016/j.neuroimage.2022.118933

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural Language Descriptions of Deep Visual Features. *ArXiv:2201.11114 [Cs]*. http://arxiv.org/abs/2201.11114

Hilbert, K., Lueken, U., Muehlhan, M., & Beesdo‐Baum, K. (2017). Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: A multimodal machine learning study. *Brain and Behavior*, *7*(3), e00633. https://doi.org/10.1002/brb3.633

Hippocampal volume and depression: A meta-analysis of MRI studies—PubMed. (n.d.). Retrieved March 26, 2022, from https://pubmed.ncbi.nlm.nih.gov/15514393/

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffer, E., & Ailon, N. (2018). Deep metric learning using Triplet network. *ArXiv:1412.6622 [Cs, Stat]*. http://arxiv.org/abs/1412.6622

How the machine 'thinks': Understanding opacity in machine learning algorithms—Jenna Burrell, 2016. (n.d.). Retrieved March 26, 2022, from https://journals.sagepub.com/doi/full/10.1177/2053951715622512

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv:1704.04861 [Cs]*. http://arxiv.org/abs/1704.04861

Huang, X., Gong, Q., Sweeney, J. A., & Biswal, B. B. (2019). Progress in psychoradiology, the clinical application of psychiatric neuroimaging. *The British Journal of Radiology*, *92*(1101), 20181000. https://doi.org/10.1259/bjr.20181000

Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning? *ArXiv:1608.08614 [Cs]*. http://arxiv.org/abs/1608.08614

Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology—PubMed. (n.d.). Retrieved January 28, 2022, from https://pubmed-ncbi-nlm-nih-gov.proxy-um.researchport.umd.edu/32530098/

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv:1502.03167 [Cs]*. http://arxiv.org/abs/1502.03167

Jafri, M. J., Pearlson, G. D., Stevens, M., & Calhoun, V. D. (2008). A Method for Functional Network Connectivity Among Spatially Independent Resting-State Components in Schizophrenia. *NeuroImage*, *39*(4), 1666–1681. https://doi.org/10.1016/j.neuroimage.2007.11.001

Jan, Z., Al-Ansari, N., Mousa, O., Abd-alrazaq, A., Ahmed, A., Alam, T., & Househ, M. (2021). The Role of Machine Learning in Diagnosing Bipolar Disorder: Scoping Review. *Jornal of Medical Internet Research*, *23*(11). https://www.jmir.org/2021/11/e29749

Javitt, D. C. (2014). Balancing therapeutic safety and efficacy to improve clinical and economic outcomes in schizophrenia: A clinical overview. *The American Journal of Managed Care*, *20*(8 Suppl), S160-165.

JMIR Medical Informatics—Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice. (n.d.). Retrieved January 28, 2022, from https://medinform.jmir.org/2021/9/e28776

Kaji, D. A., Zech, J. R., Kim, J. S., Cho, S. K., Dangayach, N. S., Costa, A. B., & Oermann, E. K. (2019). An attention based deep learning model of clinical events in the intensive care unit. *PLOS ONE*, *14*(2), e0211057. https://doi.org/10.1371/journal.pone.0211057

Kalmady, S. V., Greiner, R., Agrawal, R., Shivakumar, V., Narayanaswamy, J. C., Brown, M. R. G., Greenshaw, A. J., Dursun, S. M., & Venkatasubramanian, G. (2019). Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *Npj Schizophrenia*, *5*(1), 1–11. https://doi.org/10.1038/s41537-018-0070-8

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). The Kinetics Human Action Video Dataset. *ArXiv:1705.06950 [Cs]*. http://arxiv.org/abs/1705.06950

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual Review of Public Health*, *34*, 119–138. https://doi.org/10.1146/annurev-publhealth-031912-114409

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept

Activation Vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, 2668–2677. https://proceedings.mlr.press/v80/kim18d.html

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (n.d.). *HMDB: A Large Video Database for Human Motion Recognition*. 8.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, Z., Pan, H., Zhu, Y., & Qin, A. K. (2020). PGD-UNet: A Position-Guided Deformable Network for Simultaneous Segmentation of Organs and Tumors. *ArXiv:2007.01001 [Cs, Eess]*. http://arxiv.org/abs/2007.01001

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

Lipton, Z. C. (n.d.). In machine learning, the concept of interpretability is both important and slippery. *Machine Learning*, 28.

Lui, S., Zhou, X. J., Sweeney, J. A., & Gong, Q. (2016). Psychoradiology: The Frontier of Neuroimaging in Psychiatry. *Radiology*, *281*(2), 357–372. https://doi.org/10.1148/radiol.2016152149

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv:1705.07874 [Cs, Stat]*. http://arxiv.org/abs/1705.07874

Mao, Z., Su, Y., Xu, G., Wang, X., Huang, Y., Yue, W., Sun, L., & Xiong, N. (2019). Spatio-temporal deep learning method for ADHD fMRI classification. *Information Sciences*, *499*.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263–269. https://doi.org/10.1016/j.jalz.2011.03.005

Mentally Sick or Not—(Bio)Markers of Psychiatric Disorders Needed—PMC. (n.d.). Retrieved March 26, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7465438/

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

MRI. (2021). Mayo Clinic. https://www.mayoclinic.org/tests-procedures/mri/about/pac-20384768

MRI - Mayo Clinic. (n.d.). Retrieved March 26, 2022, from https://www.mayoclinic.org/tests-procedures/mri/about/pac-20384768

Muehlematter, U. J., Daniore, P., & Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe

(2015–20): A comparative analysis. *The Lancet Digital Health*, *3*(3), e195–e203. https://doi.org/10.1016/S2589-7500(20)30292-2

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W., & Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, *1*(1), 55–66. https://doi.org/10.1016/j.jalz.2005.06.003

Mumtaz, W., & Qayyum, A. (2019). A deep learning framework for automatic diagnosis of unipolar depression. *International Journal of Medical Informatics*, *132*, 103983. https://doi.org/10.1016/j.ijmedinf.2019.103983

Mundhenk, T. N., Chen, B. Y., & Friedland, G. (2020). Efficient Saliency Maps for Explainable AI. *ArXiv:1911.11293 [Cs]*. http://arxiv.org/abs/1911.11293

Neuroimaging in psychiatric disorders—PubMed. (n.d.). Retrieved January 28, 2022, from https://pubmed-ncbi-nlm-nih-gov.proxy-um.researchport.umd.edu/21274689/

Ni, D., Pan Chui, Y., Qu, Y., Yang, X., Qin, J., & Wong, T.-T. (2009). Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT. *Computerized Medical Imaging and Graphics*, *33*(2), 559–566.

NIMH Data Archive—Data—Collection. (n.d.-a). Retrieved March 26, 2022, from https://nda.nih.gov/edit_collection.html?QA=false&id=2126

NIMH Data Archive—Data—Collection. (n.d.-b). Retrieved March 26, 2022, from https://nda.nih.gov/edit_collection.html?QA=false&id=2274

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, *15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., & Walsh, J. (2019). *Deep Learning vs. Traditional Computer Vision*. Computer Vision Conference, Las Vegas, NV, USA. https://arxiv.org/ftp/arxiv/papers/1910/1910.13796.pdf

Owen, M. J., Sawa, A., & Mortensen, P. B. (2016). Schizophrenia. *Lancet (London, England)*, *388*(10039), 86–97. https://doi.org/10.1016/S0140-6736(15)01121-6

Paul, Y., Hickok, E., Sinha, A., Tiwari, U., & Bidare, P. M. (n.d.). *Artificial Intelligence in the Healthcare Industry in India*. 45.

Personal details of over 200,000 Malaysian organ donors leaked online: Report | Reuters. (n.d.). Retrieved January 28, 2022, from https://www.reuters.com/article/us-malaysia-cybercrime/personal-details-of-over-200000-malaysian-organ-donors-leaked-online-report-idUSKBN1FD07B

Polanczyk, G., Silva de Lima, M., Lessa Horta, B., Biederman, J., & Agusot Rohde. (2007). The Worldwide Prevalence of ADHD: A Systematic Review and Metaregression Analysis. *The American Journal of Psychiatry*. https://ajp.psychiatryonline.org/doi/full/10.1176/ajp.2007.164.6.942?casa_token=FYuLzjHNnjUAAAAA%3AdYfbA2gKXOPtaD1ajTsZcfzUSLHlBG2XOyCYCpGYgcqnhaXUCpuqpHwBy33fVWH_dGo-Q0zRsWxFpw

Pominova, M., Kondrateva, E., Sharaev, M., Pavlov, S., Bernstein, A., & Burnaev, E. (2019). 3D Deformable Convolutions for MRI classification. *ArXiv:1911.01898 [Cs, Eess]*. http://arxiv.org/abs/1911.01898

Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., & Hoffmann, H. (2019). Explainability Methods for Graph Convolutional Neural Networks. *2019 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), 10764–10773. https://doi.org/10.1109/CVPR.2019.01103

Progress in psychoradiology, the clinical application of psychiatric neuroimaging—PubMed. (n.d.). Retrieved January 28, 2022, from https://pubmed-ncbi-nlm-nih-gov.proxy-um.researchport.umd.edu/31170803/

Psychoradiology: The Frontier of Neuroimaging in Psychiatry. (n.d.). Retrieved January 28, 2022, from https://www-ncbi-nlm-nih-gov.proxy-um.researchport.umd.edu/pmc/articles/PMC5084981/

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-Alone Self-Attention in Vision Models. *ArXiv:1906.05909 [Cs]*. http://arxiv.org/abs/1906.05909

RDoC Frequently Asked Questions (FAQ). (n.d.). National Institute of Mental Health (NIMH). Retrieved March 26, 2022, from https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/resources/rdoc-frequently-asked-questions-faq

Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT - ScienceDirect. (n.d.). Retrieved March 26, 2022, from https://www.sciencedirect.com/science/article/pii/S0895611109000640

Redirect Notice. (n.d.). Retrieved January 28, 2022, from https://www.google.com/url?q=https://www.cdc.gov/nchs/data/series/sr_01/sr01_056.pdf&sa=D&source=docs&ust=1642465121309791&usg=AOvVaw2fJYo1-3j0v5HnTFtcskN6

Research Domain Criteria (RDoC). (2009). NIMH. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv:1602.04938 [Cs, Stat]*. http://arxiv.org/abs/1602.04938

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv:1505.04597 [Cs]*. http://arxiv.org/abs/1505.04597

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *ArXiv:1409.0575 [Cs]*. http://arxiv.org/abs/1409.0575

Schnyer, D. M., Clasen, P. C., Gonzalez, C., & Beevers, C. G. (2017). Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder. *Psychiatry Research. Neuroimaging*, *264*, 1–9. https://doi.org/10.1016/j.pscychresns.2017.03.003

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Si, J., Zhang, H., Zhu, L., & Chen, A. (2021). The Relationship between Overweight/Obesity and Executive Control in College Students: The Mediating Effect of BDNF and 5-HT. *Life*, *11*(4), 313. https://doi.org/10.3390/life11040313

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv:1312.6034 [Cs]*. http://arxiv.org/abs/1312.6034

"Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. (2014). 30.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *ArXiv:1212.0402 [Cs]*. http://arxiv.org/abs/1212.0402

Sowa, K., Przegalinska, A., & Ciechanowski, L. (2021). Cobots in knowledge work: Human – AI collaboration in managerial professions. *Journal of Business Research*, *125*, 135–142. https://doi.org/10.1016/j.jbusres.2020.11.038

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for Simplicity: The All Convolutional Net. *ArXiv:1412.6806 [Cs]*. http://arxiv.org/abs/1412.6806

Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., & Vaswani, A. (2021). Bottleneck Transformers for Visual Recognition. *ArXiv:2101.11605 [Cs]*. http://arxiv.org/abs/2101.11605

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (n.d.). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. 30.

Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge—PubMed. (n.d.). Retrieved March 27, 2022, from https://pubmed.ncbi.nlm.nih.gov/25652394/

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *ArXiv:1512.00567 [Cs]*. http://arxiv.org/abs/1512.00567

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Tang, G., Müller, M., Rios, A., & Sennrich, R. (2018). Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. *ArXiv:1808.08946 [Cs]*. http://arxiv.org/abs/1808.08946

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing Machine Learning Models via Prediction APIs. *ArXiv:1609.02943 [Cs, Stat]*. http://arxiv.org/abs/1609.02943

Uyulan, C., Ergüzel, T. T., Unubol, H., Cebi, M., Sayar, G. H., Nezhad Asad, M., & Tarhan, N. (2021). Major Depressive Disorder Classification Based on Different Convolutional Neural Network Models: Deep Learning Approach. *Clinical EEG and Neuroscience*, *52*(1), 38–51. https://doi.org/10.1177/1550059420916634

Varghese, T., Raghavan, S., Mathuranath, P., & Singh, N. (2014). Discrimination between Alzheimer's Disease, Mild Cognitive Impairment and Normal Aging Using ANN

Based MR Brain Image Segmentation. In *Advances in Intelligent Systems and Computing* (Vol. 247, pp. 129–136). https://doi.org/10.1007/978-3-319-02931-3_16

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *ArXiv:1710.10903 [Cs, Stat]*. http://arxiv.org/abs/1710.10903

Videbech, P., & Ravnkilde, B. (2004). Hippocampal volume and depression: A meta-analysis of MRI studies. *The American Journal of Psychiatry*. https://pubmed.ncbi.nlm.nih.gov/15514393/

Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. *Communications Medicine*, *1*(1), 1–3. https://doi.org/10.1038/s43856-021-00028-w

Wang, Y., Gong, N., & Fu, C. (2021). Major depression disorder diagnosis and analysis based on structural magnetic resonance imaging and deep learning. *Journal of Integrative Neuroscience*, *20*(4), 977–984. https://doi.org/10.31083/j.jin2004098

Waszkiewicz, N. (2020a). Mentally Sick or Not—(Bio)Markers of Psychiatric Disorders Needed. *Journal of Clinical Medicine*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7465438/

Waszkiewicz, N. (2020b). Mentally Sick or Not—(Bio)Markers of Psychiatric Disorders Needed. *Journal of Clinical Medicine*, *9*(8), 2375. https://doi.org/10.3390/jcm9082375

Wilczyńska, K., Simonienko, K., Konarzewska, B., Szajda, S., & Waszkiewicz, N. (2018). Morphological changes of the brain in mood disorders. *Psychiatria Polska*, *52*(5), 797–805. https://doi.org/10.12740/PP/89553

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2016). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *ArXiv:1502.03044 [Cs]*. http://arxiv.org/abs/1502.03044

Yang, J., Huang, X., He, Y., Xu, J., Yang, C., Xu, G., & Ni, B. (2021). Reinventing 2D Convolutions for 3D Images. *IEEE Journal of Biomedical and Health Informatics*, *25*(8), 3009–3018. https://doi.org/10.1109/JBHI.2021.3049452

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 818–833). Springer International Publishing. https://doi.org/10.1007/978-3-319-10590-1_53

Zeng, L.-L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., Chen, X., Liu, Z., Yin, H., Tan, Q., Wang, K., & Hu, D. (2018). Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI. *EBioMedicine*, *30*, 74–85. https://doi.org/10.1016/j.ebiom.2018.03.017

Zhang, Y., McAreavey, K., & Liu, W. (2022). Developing and Experimenting on Approaches to Explainability in AI Systems: ICAART 2022. *Proceedings of the 14th*

International Conference on Agents and Artificial Intelligence (ICAART'22), *2*, 518–527. https://doi.org/10.5220/0010900300003116

Zhang-James, Y., Helminen, E. C., Liu, J., Group, T. E.-A. W., Franke, B., Hoogman, M., & Faraone, S. V. (2020). *Evidence for Similar Structural Brain Anomalies in Youth and Adult Attention-Deficit/Hyperactivity Disorder: A Machine Learning Analysis* (p. 546671). bioRxiv. https://doi.org/10.1101/546671

Zhao, H., Jia, J., & Koltun, V. (2020). Exploring Self-attention for Image Recognition. *ArXiv:2004.13621 [Cs]*. http://arxiv.org/abs/2004.13621

Zhu, X., Hu, H., Lin, S., & Dai, J. (2018). Deformable ConvNets v2: More Deformable, Better Results. *ArXiv:1811.11168 [Cs]*. http://arxiv.org/abs/1811.11168