Deep Interpretability Methods for Neuroimaging

Md Mahfuzur Rahman

Follow this and additional works at: https://scholarworks.gsu.edu/cs_diss

Recommended Citation
doi: https://doi.org/10.57709/32649249

This Dissertation is brought to you for free and open access by the Department of Computer Science at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Computer Science Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
Deep Interpretability Methods for Neuroimaging

by

Md Mahfuzur Rahman

Under the Direction of Sergey Plis, PhD

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2022
Brain dynamics are highly complex and yet hold the key to understanding brain function and dysfunction. The dynamics captured by resting-state functional magnetic resonance imaging data are noisy, high-dimensional, and not readily interpretable. The typical approach of reducing this data to low-dimensional features and focusing on the most predictive features comes with strong assumptions and can miss essential aspects of the underlying dynamics. In contrast, introspection of discriminatively trained deep learning models may uncover disorder-relevant elements of the signal at the level of individual time points and spatial locations. Nevertheless, the difficulty of reliable training on high-dimensional but small-sample datasets and the unclear relevance of the resulting predictive markers prevent the widespread use of deep learning in functional neuroimaging. In this dissertation, we address these challenges by proposing a deep learning framework to learn from high-dimensional dynamical data while maintaining stable, ecologically valid interpretations. The developed model is pre-trainable and alleviates the need to collect an enormous amount of neuroimaging samples to achieve optimal training.

We also provide a quantitative validation module, Retain and Retrain (RAR), that can objectively verify the higher predictability of the dynamics learned by the model. Results successfully demonstrate that the proposed framework enables learning the fMRI dynamics directly from small data and capturing compact, stable interpretations of features predictive of function and dysfunction. We also comprehensively reviewed deep interpretability literature in the neuroimaging domain. Our analysis reveals the ongoing trend of interpretability
practices in neuroimaging studies and identifies the gaps that should be addressed for effective human-machine collaboration in this domain.

This dissertation also proposed a post hoc interpretability method, Geometrically Guided Integrated Gradients (GGIG), that leverages geometric properties of the functional space as learned by a deep learning model. With extensive experiments and quantitative validation on MNIST and ImageNet datasets, we demonstrate that GGIG outperforms integrated gradients (IG), which is considered to be a popular interpretability method in the literature. As GGIG is able to identify the contours of the discriminative regions in the input space, GGIG may be useful in various medical imaging tasks where fine-grained localization as an explanation is beneficial.

INDEX WORDS: Neuroimaging, Disorder dynamics, Deep learning, Contrastive self-supervised learning, Explainable AI, Post hoc interpretability
Deep Interpretability Methods for Neuroimaging

by

Md Mahfuzur Rahman

Committee Chair:

Sergey Plis

Committee:

Rajshekar Sunderraman

Robyn Miller

Vince Calhoun

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2022
DEDICATION

This dissertation is dedicated to my beloved wife, Sharmin, my daughter Anandita, and my father’s departed soul. Sharmin had heard and shared all the joys, struggles, and pain of my heart during the whole journey, days and nights. Anandita, like her name, was a spontaneous source of joy and inspiration. Her birth was the greatest gift I have ever received. My father’s departure last year made me realize I loved him more than I could think. His departed soul gave me a new realization of what underlies this apparent dynamism of life.
First of all, I am grateful to my parents, for I would never be here without them. I am whole-heartedly grateful to my family, especially to my wife, Sharmin, who was with me throughout my entire journey with all the patience, support, and care one could ever expect from their partner. For her presence, despite many struggles and hardships, there was joy and hope. My fatherhood at the beginning of my Ph.D. journey was a great blessing for me. Anandita, my daughter, was a continuous source of inspiration enough to obliterate all the struggles and frustrations during this journey. I miss my father here, who will never hear the news of my Ph.D. He was one of the few people in my life who would be excited to hear about this great achievement.

I want to express my gratitude to my advisor, Dr. Sergey Plis, for his continuous support, guidance, motivation, and patience throughout my Ph.D. journey. His encouragement and support significantly shaped my academic and personal growth. Without his guidance and support, it would not be possible for me to finish this endeavor. I am grateful to Dr. Rajshekhar Sunderraman for his incredibly valuable mentorship throughout this Ph.D. study. I am also thankful to other committee members—Dr. Robyn Miller and Dr. Vince Calhoun for their support and guidance toward this dissertation. I am thankful to all the teachers in my academic life whose teachings and guidance supported my growth with appropriate resonance. I am grateful to my fellow Ph.D. students and friends who significantly influenced my learning and life at GSU, especially Rashed, Khadija, Mahbub, Usman, and Noah.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ................................................................. v

LIST OF TABLES ................................................................. x

LIST OF FIGURES ............................................................... xi

LIST OF ABBREVIATIONS .................................................... xvi

1 INTRODUCTION ................................................................. 1

1.1 Background ................................................................. 1

1.2 Challenges in Deep Learning Approach to Neuroimaging .......... 3

1.3 Aims of This Dissertation ................................................ 4

1.4 Dissertation Contributions ............................................. 7

1.5 Dissertation Outline ..................................................... 11

2 EXPLAINABLE AI: A LITERATURE REVIEW ............................... 13

2.1 What and Why Is Model Interpretability? ............................. 13

2.2 Philosophy of Scientific Explanations .................................. 15

2.2.1 How to Achieve Interpretability? .................................... 16

2.2.2 Global vs. Local Interpretability ................................. 16

2.3 Taxonomy of Model Interpretability Problem ......................... 17

2.4 Important Terminology in Interpretability ............................ 18

2.5 Axioms of Attribution Methods ........................................ 20

2.6 Taxonomy of Interpretability Methods .................................. 22

2.7 Visualization Methods .................................................... 25

2.7.1 Gradient-Based Methods ............................................ 25

2.7.2 Perturbation-Based Methods: ...................................... 32
2.8 Distillation Methods ........................................ 33
2.9 Intrinsic Methods ........................................... 36
   2.9.1 Attention Mechanism .................................. 36
   2.9.2 Joint Training ........................................ 37
   2.9.3 Modular Transparency ................................ 37
2.10 Counterfactual Explanations ............................... 38
2.11 Influence Functions ...................................... 39
2.12 Sanity Checks for Interpretability Methods ............ 39
2.13 Evaluation Metrics ....................................... 40
   2.13.1 Metrics for Ground-truth Datasets .................. 41
   2.13.2 Metrics for Real Datasets ............................ 41
2.14 Criticisms of Post hoc Interpretability ................. 46
2.15 Conclusion ................................................. 47

3 DEEP LEARNING INTERPRETABILITY IN NEUROIMAGING  . . 48
3.1 Introduction ................................................. 48
   3.1.1 Feature Engineering Approach to Neuroimaging ...... 49
   3.1.2 Deep Learning Approach to Neuroimaging ............ 51
3.2 Related Work ............................................... 53
3.3 A Quick Reference to Interpretability in Neuroimaging .... 54
3.4 Transfer Learning in Neuroimaging ........................ 57
3.5 Review of Interpretability Methods in Neuroimaging .... 58
   3.5.1 Gradient-based Methods ............................... 58
   3.5.2 Perturbation-based Methods ............................ 67
   3.5.3 Distillation Methods .................................. 68
   3.5.4 Feature Map Visualization ............................. 74
   3.5.5 Intrinsic Methods ..................................... 75
3.6 The Usage Trend of Interpretability Methods ............ 75
3.7 Suggestions for Interpretable Models in Neuroimaging ............ 79
3.8 Conclusion ........................................................................... 86

4 TRANSFER LEARNING FOR NEUROIMAGING ......................... 87
4.1 Introduction .......................................................................... 87
4.2 Related Work ......................................................................... 88
4.3 Method Description .............................................................. 89
4.4 Experiments ................................................................. 90
   4.4.1 Simulation Data .......................................................... 90
   4.4.2 Real data ..................................................................... 91
   4.4.3 Training ...................................................................... 92
   4.4.4 Results ...................................................................... 93
4.5 Conclusions and Future Work ............................................. 94

5 INTERPRETING BRAIN DYNAMICS .......................................... 96
5.1 Introduction .......................................................................... 96
5.2 Results .................................................................................. 99
   5.2.1 Datasets .................................................................... 101
   5.2.2 whole MILC Performance ........................................... 101
   5.2.3 whole MILC Evaluation ............................................. 101
   5.2.4 Post hoc Explanation and RAR Evaluation ................. 103
5.3 Discussion ........................................................................... 111
5.4 Methods ................................................................................ 115
   5.4.1 Preprocessing .......................................................... 117
   5.4.2 Whole MILC ............................................................. 118
   5.4.3 Attention Mechanism .............................................. 121
   5.4.4 whole MILC Setup .................................................. 122
   5.4.5 Model Interpretability ............................................... 123
   5.4.6 Random Baseline ..................................................... 124
5.4.7 RAR Method and Setup ........................................... 125
5.5 Conclusion ............................................................... 126

6 GEOMETRICALLY GUIDED INTEGRATED GRADIENTS .......... 128

6.1 Introduction ............................................................. 128
6.2 Related Work ............................................................ 131
6.3 Geometrically Guided Integrated Gradients ...................... 133
6.4 Model Perturbation Sanity Check .................................... 136
6.5 Experiments ............................................................... 137
  6.5.1 MNIST Experiments ................................................ 138
  6.5.2 Quantitative Evaluation on MNIST Dataset .................. 140
  6.5.3 Experiments on ImageNet .......................................... 141
  6.5.4 Target Object and Model Parameter Sensitivity Tests ...... 142
  6.5.5 Evaluation of Attribution Quality ............................... 144
6.6 Conclusion ............................................................... 149

7 CONCLUSION AND FUTURE WORK ................................. 150

7.1 Summary ................................................................. 150
7.2 Future Directions ....................................................... 153
Appendices ................................................................. 157
A Complete RAR Evaluation Results ................................. 158

REFERENCES ............................................................... 159
LIST OF TABLES

Table 3.1 Illustrative neuroimaging studies using popular interpretability methods 56
Table 3.2 Literature Review of Interpretable Deep Learning Research in Neuroimaging 69
Table 3.2 Literature Review of Interpretable Deep Learning Research in Neuroimaging 70
Table 3.2 Literature Review of Interpretable Deep Learning Research in Neuroimaging 71
Table 3.2 Literature Review of Interpretable Deep Learning Research in Neuroimaging 72
Table 3.2 Literature Review of Interpretable Deep Learning Research in Neuroimaging 73
Table 3.3 The usage trend of popular post hoc interpretability methods. . . . . . . 76
Table 3.3 The usage trend of popular post hoc interpretability methods. . . . . . . 77
Table 3.3 The usage trend of popular post hoc interpretability methods. . . . . . . 78
Table 3.3 The usage trend of popular post hoc interpretability methods. . . . . . . 79
Table 6.1 Summary of Quantitative Evaluation on MNIST dataset . . . . . . . . . . . 140
Table 6.2 Summary of SIC and AIC Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . 148
LIST OF FIGURES

Figure 1.1 An overview of our approach to model interpretation. A: Construct a model for disorder-specific discovery: the DL model learns directly from the disorder signal dynamics and retains interpretations for further introspection. B: Leverage self-supervised pretraining to distinguish healthy subjects: Learned representations assist the model in maintaining its predictive power when downstream training data is limited. C: Construct a downstream model to discriminate patients from controls for each disorder, starting with the pre-trained weights: transfer of representations learned during pretraining simplifies convergence and balances overfitting. D: Introspection of the trained downstream models: interpretability methods extract meaningful, distinctive parts through feature attributions. Subsequently, the estimated salient aspects of the dynamics go through an automatic validation process. This information can then be relayed to a human expert in the relevant field to interpret further and advance knowledge about the disorders. .............................................. 6

Figure 2.1 Taxonomy of Explainable AI in terms of transparency and holistic view. 18

Figure 2.2 Taxonomy of Explainable AI. We also show how frequently these methods have been used in neuroimaging studies. ................................. 24

Figure 3.1 Left: "Explainable AI" is getting popular or becoming an area of concern over the years (2012 - 2022) as reflected in the Google Trends Popularity Index (Max. value is 100). Right: To get relevant statistics, we searched with the keywords "deep learning in neuroimaging" and "interpretability in deep learning" at this website https://app.dimensions.ai/discover/publication (Accessed on October 13, 2022). Neuroimaging studies increasingly used deep learning models during the last decade (2012 - 2021) to understand the dynamics of brain functions and anatomical structures. The need to interpret black-box models is growing accordingly. ................................. 55

Figure 4.1 Left: Classification performance of schizophrenia patients from normal controls on the progressively larger datasets. Right: Simulated VAR vs. SVAR time-series classification performance. As we can observe, the ST-DIM pretraining improved the downstream discriminative performance for both the real and synthetic experiments. ................................. 93
Figure 5.1  An overview of our approach to model interpretation (created in program Inkscape 1.1.2, inkscape.org/release/inkscape-1.1.2). **A:** Construct a model for disorder-specific discovery: we divided the entire ICA time courses into multiple sliding windows. Then we fed them into the whole MILC model that learns directly from the disorder signal dynamics and retains interpretations for further introspection. **B:** Leverage self-supervised pretraining to distinguish healthy subjects: learned representations assist the model in maintaining its predictive power when downstream training data is limited. **C:** Construct a downstream model to discriminate patients from controls for each disorder starting with the pre-trained whole MILC weights: transfer of representations learned during pretraining simplifies convergence and balances overfitting. **D:** Introspection of the trained downstream models: we compute saliency maps as a rationale used by the model behind every prediction using interpretability methods to extract meaningful, distinctive parts of the data. Subsequently, the estimated salient aspects of the dynamics go through an automatic validation process. To this end, we use the most salient features to retrain an independent SML model that confirms the salience of the features. This information can then be relayed to a human expert in the relevant field to interpret further and advance knowledge about the disorders. **E:** Examples of saliency maps as deemed highly predictive by the models for their predictions in three different discriminative tasks. Please note that the red boxes mark the highly discriminative salient parts of the data.

Figure 5.2  The main results from the whole MILC architecture and its comparison with standard machine learning models (SML). Apparently, the whole MILC model, in general, can learn from the raw data where traditional SML models fail to maintain their predictive capacity. Moreover, the whole MILC w/ pretraining substantially improves the latent representations as reflected in the improved accuracy compared to the whole MILC w/o pretraining. Specifically, in most small data cases, the whole MILC w/ pretraining outperformed the whole MILC w/o pretraining across the datasets. However, as expected, when we gradually increased the number of subjects during training, the effect of pretraining on the classification performance diminished, and both configurations of whole MILC did equally well. We verified this trend over three datasets that correspond to autism spectrum disorder, schizophrenia, and Alzheimer’s disease. Please note the Wilcoxon rank test results between w/ and w/o pretraining performance of the model as marked by asterisk(∗) and "ns" (not significant), where ns : p > 5e⁻², ∗ : 1e⁻² < p ≤ 5e⁻², ∗∗ : 1e⁻³ < p ≤ 1e⁻², ∗∗∗ : 1e⁻⁴ < p ≤ 1e⁻³, ∗∗∗∗ : p ≤ 1e⁻⁴.
RAR employs SVM to classify the FNCs of the top 5% of the salient input data as estimated by the whole MILC model’s predictions. We used integrated gradients (IG) and smoothgrad integrated gradients (SGIG) to compute feature attributions. It is evident that when an independent classifier (SVM) learned on every subject’s most salient 5% data, the predictive power was significantly higher compared to the same SVM model trained on the randomly chosen same amount of data. In other words, the poor performance with randomly selected data parts indicates that other parts of the data were not exclusively discriminative as the whole MILC estimated salient 5% data parts. We also notice that sample masks over a different percentage of data coverage gradually obscured the localization of the discriminative activity within the data. Though the SVM model gradually became predictive with increased randomly selected data coverage, which we show in Supplementary Information, this performance upgrade was due to the gradual improvement in functional connectivity estimation and not attributable to the disease-specific localized parts within the data. For every disorder (Autism spectrum disorder, Schizophrenia, and Alzheimer’s disease), the higher AUC at this 5% indicates stronger relevance of the salient data parts to the underlying disorders. Furthermore, the RAR results reflect that in most cases, when whole MILC was trained with limited data, the w/ pretraining models estimated feature attributions more accurately than the models w/o pretraining.

Top 10% FNC for patients computed using most 5% of the salient data as thresholded using feature attribution maps (saliency maps) for different disorders (created in programs MNE 1.1.dev0, https://mne.tools/dev/ and Inkscape 1.1.2, inkscape.org/release/inkscape-1.1.2). Apart from the high predictive capacity of the salient data, we observed some intriguing differences among these connectograms. The autism spectrum disorder exhibits the lowest between-domain FNC. However, salient data in autism disorder highlights domain changes in specific cerebellum, sensorimotor, and subcortical domains. The model-identified salient data reflects the most widespread pattern for schizophrenia and is consistent with the literature showing cerebellum interaction across multiple domains and sensorimotor changes. The predictive features for Alzheimer’s disease mainly concentrate on visual and cognitive interactions.
Figure 5.5  **A:** Full FNC for patients computed using most 5% of the salient data selected based on feature attribution values for different disorders. **B:** Static FNC (i.e., using 100% data) matrices for patients of different disorders. The FNC based on 5% salient data (**A**) does indeed convey the same focused dynamic information as currently assessed in FNC matrices based on 100% data (**B**). It is thus apparent that the proposed model can capture the focused information aligned with the current domain knowledge. **C:** Pairwise difference of FNC matrices based on 5% salient data. The difference FNC matrices based on focused data indicate that each disorder has a uniquely distinguishable association with brain dynamics.

Figure 5.6  **A:** Examples of the temporal density based on the top 5% values of the saliency maps from patients and controls for each disorder. It is noticeable that the temporal density for schizophrenia and Alzheimer’s patients is more focal in time as reflected in the spikiness, indicating that the discriminative activity for patients occurs predominantly in a shorter time interval. In contrast, for controls, model predictions do not relate to specific time intervals. For autism spectrum disorder, however, the whole MILC model did not capture any temporal adherence to the discriminative activity for patients. That is, the discriminatory events are not focal on shorter time intervals for ASD. **B:** The EMD (Earth Mover’s Distance) distributions as a proxy measure for uniformity/spikiness of temporal densities (edited in program Inkscape 0.92.2, inkscape.org/release/0.92.2/). We analyzed the EMD measures of patients and controls to investigate the discriminative properties of salient data in terms of the spikiness or uniformity of the temporal densities. The larger EMD measures for schizophrenia and Alzheimer’s patients substantiate that the model found the discriminative activity in shorter focused time intervals. In contrast, for ASD, the equal EMD values for both patients and controls indicate that the temporal density measures do not relate to the discriminative activity for this disorder. We verified these observations with the statistical significance (Wilcoxon rank) test results as marked by asterisk(*) and "ns" (not significant), where ns : $p > 5e^{-2}$, **** : $p \leq 1e^{-4}$. 
Figure 5.7  The whole MILC architecture—an attention-based top-down recurrent network (created in programs Adobe Illustrator 26.0.3 and Inkscape 1.1.2). Precisely, we used an LSTM network with an attention mechanism as a parameter-shared encoder to generate the latent embeddings $z$ for the sliding window at all relevant positions. The top LSTM network (marked as LSTM) used these embeddings ($z$) to obtain the global representation $c$ for the entire subject. During pretraining, we intended to maximize the mutual information between $z$ and $c$. In the downstream classification task, we used the global representation $c$ directly as input to a fully connected network for predictions. Based on these predictions, we estimated feature attributions using different interpretability methods. Finally, we evaluated the feature attributions using the RAR method and an SVM model.

Figure 5.8  End-to-end process of RAR evaluation. For each subject in the dataset, based on the whole MILC class prediction and model parameters, we estimated the feature importance vector $e$ using some interpretability method $g_i$. Later on, we validated these estimates against random feature attributions $g_R$ using the RAR method and an SVM model. Through the SVM model’s performance when separately trained with different feature sets, we show that whole MILC model-estimated features were highly predictive compared to a random selection of a similar amount of features. Empirically, we show that $\xi(X^M | g_i) > \xi(X^M | g_R)$, where $\xi$ is the performance evaluation function (e.g., area under the curve) and $X^M$ refers to the modified dataset constructed based on only retained feature values.

Figure 6.1  Comparison of gradients, integrated gradients and GGIG based on top 5% pixel attributions for the "gazelle" prediction. In GGIG explanation, the details of the animal are clearly noticeable and discarded the undesirable regions from the explanation.

Figure 6.2  Overall pipeline of the work. First, we generate explanations for each prediction. Secondly, for quantitative evaluation, we start with a complete defocused image and combine with only salient pixels (thresholded at $x\%$) to create a test image. Finally, we feed the model with the test image and measure both accuracy and relative softmax scores.

Figure 6.3  The working mechanism of GGIG. It takes a baseline $x'$, creates a linearly spaced path to the actual input $x$. From each interpolation point, it follows a trajectory in the loss landscape that maximizes class activation.
Figure 6.4  **Left:** Comparison of saliency-focused images thresholded at 10% while we perturb the model. **Right:** Relative softmax score during model perturbation. As expected, the similarity with the original explanations breaks and softmax score drops with gradual increase in perturbation level.

Figure 6.5  **Top:** Selected maps for MNIST samples generated using different methods. **Bottom:** Quantitative evaluations using different correlation and similarity metrics. It is obvious that the maps produced by GGIG have higher structural and numerical correlations with the input. Precisely, GGIG, as an explanation method, performs the best by a large quantitative margin.

Figure 6.6  Sample saliency maps for ImageNet Validation set and ResNet101 model generated by gradients, integrated gradients, geometrically guided integrated gradients, and Edge Detector.

Figure 6.7  Comparative saliency maps for Inception V3 model and ImageNet dataset generated by GRAD, IG, GGIG, and Edge Detector. The maps obtained using GGIG are more discriminative and more clearly reveal the underlying structure of the class-associated objects.

Figure 6.8  The input image is from ImageNet: we show the input image and 10% salient pixels of the explanations (saliency maps) generated using 'GRAD', 'IG' and 'GGIG' (our method). As we can see, while GRAD and IG are noisy and not focusing on any discriminative parts of the image, GGIG is demonstrably focusing on very distinguishable regions which are representative of a gentleman. For example, salient predictive regions include facial parts, tie, collar, wrist watch etc, which is highly aligned to human understanding of the input image class.

Figure 6.9  "Target Object Sensitivity" test experiments on ImageNet samples and Inception V3 model. While the traditional edge detector supposedly identifies all of the sharp changes and does not focus any specific attention to the actual objects, the saliency methods are still focusing on the actual objects for predictions.

Figure 6.10  "Model Parameter Sensitivity" test experiments on ImageNet samples and Inception V3. We show how different methods generate saliency maps when layers of the underlying model (Inception V3) are gradually randomized from top to bottom.

Figure 6.11 Comparison of saliency-focused images thresholded at 10%. GGIG, compared to other methods, is more convincing from the human standpoint because it directly points to the discriminative parts of the image while ignoring redundant and unexpected regions.
Figure 6.12 **Left:** Evaluation of saliency methods using Softmax Information Curve (SIC) metric for Inception V3 and ResNet101 models using ImageNet test and validation images respectively **Right:** Evaluation using Accuracy Information Curve (AIC) metric for the same models and datasets. . . . . . . . . . . . . 148

Figure 1 RAR employs an SVM to validate FNCs computed using different percentages of the salient input data (5% - 30%) as determined through post hoc explanations of the whole MILC model’s predictions. Salient features are highly predictive compared to the same amount of random features. . . . . . . 158
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>SML</td>
<td>Standard Machine Learning</td>
</tr>
<tr>
<td>DL</td>
<td>Deep Learning</td>
</tr>
<tr>
<td>IG</td>
<td>Integrated Gradients</td>
</tr>
<tr>
<td>XRAI</td>
<td>eXplanation with Ranked Area Integrals</td>
</tr>
<tr>
<td>GGIG</td>
<td>Geometrically Guided Integrated Gradients</td>
</tr>
<tr>
<td>Grad-CAM</td>
<td>Gradient-weighted Class Activation Map</td>
</tr>
<tr>
<td>Guided-BP</td>
<td>Guided Backpropagation</td>
</tr>
<tr>
<td>LIME</td>
<td>Local Interpretable Model-agnostic Explanations</td>
</tr>
<tr>
<td>SHAP</td>
<td>SHapley Additive exPlanation</td>
</tr>
<tr>
<td>DeepLIFT</td>
<td>Deep Learning Important FeaTures</td>
</tr>
<tr>
<td>LRP</td>
<td>Layer-wise Relevance Propagation</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long-short term memory</td>
</tr>
<tr>
<td>ROAR</td>
<td>Remove and Retrain</td>
</tr>
<tr>
<td>RAR</td>
<td>Retain And Retrain</td>
</tr>
<tr>
<td>ROAD</td>
<td>RemOve And Debias</td>
</tr>
<tr>
<td>SZ</td>
<td>Schizophrenia</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer’s Disease</td>
</tr>
<tr>
<td>ASD</td>
<td>Autism Spectrum Disorder</td>
</tr>
<tr>
<td>ABIDE</td>
<td>Autism Brain Imaging Data Exchange</td>
</tr>
<tr>
<td>FBIRN</td>
<td>Function Biomedical Informatics Research Network</td>
</tr>
<tr>
<td>OASIS</td>
<td>Open Access Series of Imaging Studies</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td><strong>FNC</strong></td>
<td>Functional Network Connectivity</td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

Advancing our understanding of brain dynamics is the underpinning to uncovering the underlying neurological conditions [1–3]. Thus, localization and interpretation of subject-specific spatial and temporal activity may help guide our understanding of the disorder. Functional magnetic resonance imaging (fMRI) can noninvasively capture the brain dynamics [4, 5]. However, the complex fMRI dynamics are not directly interpretable due to the excessive dimensionality, commonly referred to as the curse of dimensionality problem [6]. A persistent goal of Artificial Intelligence (AI) in neuroimaging domain is leveraging magnetic resonance imaging (MRI) data to enable machines to learn the functional dynamics or anatomical alterations associated with underlying neurological disorders.

1.1 Background

Over the last few decades, for brain solutions, researchers have traditionally relied on analytical techniques that consider each voxel (3D pixel) of the brain independently to find what makes patients different from non-patient counterparts [7]. While this approach is promising to some extent and has helped advance our understanding of brain functions, it has some significant limitations [6, 8]. In particular, it assumes that voxels are independent and that scientists draw conclusions based on multiple independent statistical analyses. However, the current understanding of brain functions and structure reveals that the changes in different brain networks can best explain brain disorders [9, 10]. Moreover, a brain network is not necessarily spatially localized. Another limitation of this non-AI approach is that it can deal
with group-level analysis rather than individual-level decision-making. However, to receive the full translational impact of these studies on clinical practices, clinicians must deal with each individual as a separate case. These limitations naturally encouraged people to look for AI-led brain solutions, specifically the use of machine learning (ML) in understanding mental disorders [11–13]. Instead of looking into the brain regions independently, ML models look for undiscovered holistic patterns from the data using the advanced knowledge of applied statistics and mathematical optimization techniques [14].

Moreover, ML can generate individual-level diagnostic and prognostic decisions. Along these lines, standard machine learning (SML) models gained a varying degree of success, and the expert-led feature extraction and selection step is almost a prerequisite for its well-functioning [15]. However, these representations heavily rely on strong assumptions and can miss essential aspects of the underlying dynamics. Unfortunately, when trained on raw data, SML models did not perform well [16–18]. However, we need to go beyond existing knowledge, and learning from the raw data is essential for further advancement in mental health analysis. Specifically, direct learning from the data may reveal undiscovered and valuable patterns within the data and may bring translational value to clinical practices. It may also accelerate diagnostic and prognostic decision-making processes, eventually leading to personalized treatment plans. While SML models fail to learn from the raw data, deep learning (DL) has been very popular because it does not require prior feature selection or intermediate intervention [19–23]. It can learn automatically from the raw data and find discriminative and potentially useful clinical features.
1.2 Challenges in Deep Learning Approach to Neuroimaging

Transitioning from the ML approach to the DL approach demands a shift from a “feature engineering" to a “feature learning" paradigm to achieve better predictive performance and desired model introspection opportunities. Numerous DL models witnessed superior predictive performances in neuroimaging studies [8, 16, 24, 25]. Nonetheless, DL approaches in neuroimaging undergo several challenges:

One of the major concerns in neuroimaging studies is the lack of sufficient training samples [26, 27], which is hostile to the efficient training of DL models [28]. This constraint is due to the expensive data collection process in neuroimaging studies [29]. In such a scenario, transfer learning can be a convenient approach to deal with this problem, as reported in several studies [30–34]. Adapting transfer learning in neuroimaging domain is a harder problem due to the unavailability of transferable tasks and lack of ground truth. Several research studies from medical fields have widely used transfer learning from the natural imaging domain to different neuroimaging [35] or other medical imaging domains. However, Raghu et al. [36] argued the acceptability and utility of such knowledge transfer. So, formulating a suitable task that supports transferrable representation learning from unrelated neuroimaging datasets is essential to support studies dealing with limited training data.

While the interpretation of DL models may faster uncover domain-specific knowledge [37, 38], deep learning models are black-boxes [39], and there are opinions against attempts to explain black-boxes [40, 41] because of the associated risks of the unfaithfulness of the post hoc explanations. While most of the existing saliency methods highly depend on heuristics
and are constrained to suitable model architectures [42] that support robust and stable sensitivity analyses [43, 44], investigating the possibility of their usefulness is a challenging but essential step before applying them. Moreover, there is no agreed validation method for the post hoc explanations in neuroimaging studies, hindering the widespread use of automatic discovery.

Lastly, the existing interpretability methods focus on different aspects of the model’s learned behavior [45, 46]. Many methods are blamed to be computationally expensive [47, 48], unstable [49], model insensitive [50, 51], noisy [43, 52, 53]. Furthermore, some methods [54, 55] are criticized for not satisfying the desirable implementation invariance [43] property. So, it is essential to improve the existing methods or propose an entirely new method that resolves some limitations. Incorporating the geometric behavior of the function space may improve interpretability. However, it poses severe challenges, and to our knowledge, no interpretability method leveraged the geometric behavior of the function as learned by a model.

1.3 Aims of This Dissertation

Deep learning has attained significant attention in diverse areas of science, technology, and engineering applications [23]. Encouraged by its great success and the potential of growing interpretability research [56, 57] to understand the mechanisms behind DL’s success, we aim to apply this tool to guide our intuition to improve our understanding of the brain dynamics in connection with different nervous system disorders.

In the following, we list the main objectives of the proposed dissertation:
Objective 1: Develop robust DL models for interpretable representation learning for neuroimaging domain. The key research agenda for this objective are:

- Can we propose a DL model that is effective for interpretable representation learning, working directly on the minimally processed data? In other words, how can we design the DL model that is able to retain the representation and is beneficial for model introspection?

- As efficient training of DL requires a large amount of data, how can we support neuroimaging studies with minimal training data?

- Can we formulate a transfer learning task that supports knowledge transfer from available unrelated and unlabeled datasets like HCP (Human Connectome Project) to multiple downstream tasks and will improve discriminative performance and interpretability?

Objective 2: Discover the existing interpretability methods suitable for interpreting fMRI dynamics. The main research agenda for this objective are:

- Investigate the existing interpretability methods to discover if some fit for neuroimaging studies. Use synthetic datasets for initial investigation.

- Apply the well-fit interpretability methods to different downstream models and generate explanations of the model predictions based on the learned dynamics.

- Propose and implement a validation approach that is able to quantify the predictability of the generated explanations.
A visual depiction of the proposed framework focusing on **Objective 1** and **Objective 2** is shown in Figure 1.1.

**Objective 3: Develop optimization-based technique(s) for post hoc local explanations.** The main research agenda for this objective are as follows:

- Can we leverage the geometric behavior of the functional space as learned by a model to improve post hoc interpretability?

- Perform sanity checks for the new post hoc interpretability method.

- Demonstrate the method’s higher predictability using different evaluation metrics.

Figure 1.1: An overview of our approach to model interpretation. **A:** Construct a model for disorder-specific discovery: the DL model learns directly from the disorder signal dynamics and retains interpretations for further introspection. **B:** Leverage self-supervised pretraining to distinguish healthy subjects: Learned representations assist the model in maintaining its predictive power when downstream training data is limited. **C:** Construct a downstream model to discriminate patients from controls for each disorder, starting with the pre-trained weights: transfer of representations learned during pretraining simplifies convergence and balances overfitting. **D:** Introspection of the trained downstream models: interpretability methods extract meaningful, distinctive parts through feature attributions. Subsequently, the estimated salient aspects of the dynamics go through an automatic validation process. This information can then be relayed to a human expert in the relevant field to interpret further and advance knowledge about the disorders.
1.4 Dissertation Contributions

While neuroimaging studies have witnessed a growing usage of DL approaches for improved representation learning of the brain dynamics [16, 25], the studies severely suffer from the data scarcity problem. Moreover, the application of transfer learning is problematic because formulating a suitable task effective for learning is not easy. Additionally, interpretability methods from the existing literature have limitations, and most of the methods may not be useful for the neuroimaging model interpretability. With all these limitations in mind, we have made the following contributions to this dissertation:

**Contributions to Objective 1:** Developing a robust end-to-end DL framework for improved representation learning and interpretability in neuroimaging

- The application of transfer learning is seemingly tricky in neuroimaging studies. To this end, we empirically show that it is possible to formulate a discriminative task in a self-supervised fashion to better represent the underlying dynamics from a large publicly available unlabeled dataset. We proposed two contrastive self-supervised pretraining methods useful to obtain a directly transferable better representation of the neuroimaging signals. The first approach, called Spatio-Temporal Deep InfoMax, improved downstream performance for a VAR vs. SVAR classification and for classifying patients from healthy controls for schizophrenia disorder. The latter approach, called whole MILC, produced significantly better downstream results in three different disorders with diverse age groups and disorders.
• We investigated different architectural tweaks, and the proposed whole MILC model is able to retain the spatiotemporal signature in the underlying signal, which is highly discriminative and informative of the underlying brain dynamics.

• We produced the following publications from these contributions:


Contributions to Objective 2: Investigation, utilization, and validation of post hoc interpretability methods for Neuroimaging

• We investigated the existing interpretability methods to discover if they would be useful for neuroimaging studies. We used a synthetic dataset first to see their behavior and identified the utility of some of the methods for neuroimaging research.

• We successfully applied the best methods to generate explanations of predictions made by all the downstream models designed for three disorder classification tasks.
• We proposed a novel quantitative validation approach, called "Retain and Retrain (RAR)", suitable for neuroimaging studies, that supports quantitative justification of the predictability of the salient regions in the explanations.

• We have developed a metric to measure if the explanation is temporally local or global (distributed). For this metric, we took advantage of Wasserstein $p$-distance. We used $p = 1$, and thus our metric reduces to Earth Mover’s Distance (EMD). However, we computed EMD from the uniform distribution, thus characterizing the local or global temporal span of the explanations.

• We did a comprehensive survey of more than 300 neuroimaging papers using interpretable DL approaches. We provide detailed analyses of the usage trends of the most popular methods and identify the gaps in the current interpretability practices in neuroimaging.

• These contributions led to the following publications:


Contributions to Objective 3: Development of an interpretability method leveraging geometric behavior of the learned function.

- We developed a post hoc interpretability method for DL models. Our method, "Geometrically Guided Integrated Gradients (GGIG)," provides fine-grained explanations compared to a popular method called "integrated gradients (IG)" [43].

- We proposed two tests, called "\( \sigma \) – perturbation" and "target object sensitivity," as part of the sanity checks for interpretability methods.

- We also demonstrated that GGIG passed the model parameter cascaded randomization test, which is another requirement that an interpretability method should satisfy.

- We provide quantitative evaluation for GGIG using some popular evaluation metrics [58]. With extensive experiments on MNIST and ImageNet datasets and multiple DL architectures, we show that GGIG outperforms IG by a large margin. Furthermore, our results show that GGIG is robust regardless of architecture and datasets.

- GGIG localizes the contours of the discriminative regions in finer detail. Hence, GGIG may be useful for various medical imaging tasks where the localization of salient regions as an explanation may be useful to better understand or deal with the disease.

- We produced the following publications from these contributions:

Apart from the previous contributions, the author has also contributed to the following publications during his doctoral study, which is not part of this dissertation:


1.5 Dissertation Outline

The remaining part of the dissertation is organized as follows: **Chapter 2** provides a comprehensive review of the interpretability/explainability practices in AI, their philosophical contexts, interpretability desiderata, interpretability methods, and evaluation metrics. In **Chapter 3**, we introduce some necessary background about the interpretability practices in neuroimaging studies. We comprehensively review recent interpretable DL approaches in the brain imaging domain. We further show our analyses of the current usage trends of different interpretability methods in neuroimaging studies that use ML/DL and describe our advice based on gaps in existing approaches. In **Chapter 4**, we discuss our first contrastive self-supervised method, which is based on a spatio-temporal objective to pretrain a deep learning model. In **Chapter 5**, we propose a suitable interpretable DL framework, which is based
on our second pretraining approach called whole MILC. The proposed framework supports pretraining from unlabeled healthy subjects using a contrastive self-supervised approach. We also discuss how we investigated the learned dynamics using post hoc interpretability methods. We further discuss our "Retain and Retrain (RAR)" validation approach and how we leveraged this RAR framework to validate the generated explanations quantitatively for three downstream classification tasks. We show the successful application of the end-to-end (learning-interpreting-validating) framework across several studies covering diverse age groups and disorders. We also discuss our metric (based on Wasserstein $p$-distance with $p = 1$, essentially called EMD) to analyze the temporal span of the explanations. In Chapter 6, we introduce a novel interpretability method that leverages the geometric properties of the functional space as learned by a deep learning model. We also discuss our proposed two sanity tests, called "$\sigma$ – perturbation" and "target object sensitivity." We then discuss the performance of our interpretability method via experiments on different architectures and datasets. In Chapter 7, we summarize the dissertation and shed some light on the opportunities for future work toward interpretable neuroimaging.
CHAPTER 2
EXPLAINABLE AI: A LITERATURE REVIEW

As the field of ML is thriving rapidly, interpretability in machine learning has been an essential deciding factor before deploying the models in real-world practices. Explainability has been essential to ensure trust and prevent the models from unwanted behavior in safety-critical domains such as healthcare, medicine, finance, and law enforcement. While the field of ML is gaining maturity in recent years, ML research has begun since the second half of the 20th century [59]. In this chapter, we discuss the perspectives, axioms, and philosophical basis of interpretability in AI. We also provide a taxonomy of the interpretability methods based on current practices and briefly introduce the methods and metrics commonly used for synthetic and real datasets.

2.1 What and Why Is Model Interpretability?

ML systems, generally optimized to exhibit task performance, outperform humans on different computer vision and language processing tasks. However, the deployment of these systems requires satisfying other auxiliary desiderata such as safety, nondiscrimination, justice, and providing the right to explanation [60]. The unique purpose of model interpretability is to satisfy these additional criteria.

Traditionally, an ML system optimizes an objective function upon which it exhibits its predictive performance. However, a mere objective function does not include other desiderata of ML systems for its wide-ranging real-world scenarios. Thus, regardless of an ML system’s performance, those systems are still incomplete. In other words, stakeholders might seek
trust, causality, transferability, informativeness, and fairness as defined in [61]. Hence, as argued in [60], interpretability or, in other words, explanations can be one of many ways to make these gaps in problem visualization more evident to us. Some scenarios Doshi-Velez and Kim include:

- **Scientific Understanding/Data Interpretation:** We may want to create knowledge from an ML system. Explanations may be one of the ways to create knowledge from the machine’s learned behavior.

- **Safety:** Incorporating all the accompanying scenarios in developing an artificial agent is not feasible. In that case, an explanation may flag undesirable model behavior.

- **Ethics:** In problem formulation, one might not consider apriori to remove any potential bias, but the model may learn some unwanted discriminating pattern within the data.

- **Mismatched Objectives:** Often, for building an agent, one may optimize for a proxy function rather than the actual goal. In that case, the agent may discard all other factors that were very relevant to the ultimate goal. For example, a scientist may want to investigate different progressive stages of Alzheimer’s but end up building a classifier for Alzheimer’s patients from healthy controls.

- **Multi-objective Trade-offs:** When an ML system has multiple competing objectives to be satisfied, it may only be possible to incorporate some of them due to the unknown dynamics of their trade-offs.
2.2 Philosophy of Scientific Explanations

Hempel and Oppenheim (1948) [62] believed that explanation and prediction have the same logical structure, and hence they referred to explanations as "deductive systematization." Bechtel and Abrahamsen (2005) [63] viewed explanations as a mechanistic alternative and may depart from widely accepted nomological explanations, which means a phenomenon if explained, must subsume under a law. The authors deemed explanations in life sciences as "identifying the mechanism responsible for a given phenomenon." Lewis (1986) [64] viewed it as "to explain an event is to provide some information about its causal history." However, Lewis did not provide any restricted notion of what information qualifies as part of the causal history. Still, there is no formal definition of "Explainability" or "Interpretability" in the field of Artificial Intelligence [60, 61, 65]. As many researchers indicated, the ongoing interpretability practices use only researchers’ intuition that is susceptible to cognitive biases [66] and social expectations [67]. However, as de Graaf and Malle [68] hypothesized, this is not unnatural because as long as people build intentional agents, people will expect explanations from the models using the same conceptual framework people use to explain human behavior. In the current practices of "Explainable AI," the communication gap between the researchers and practitioners is evident, and Miller et al. [69] describes this phenomenon as "the inmates running the asylum." While we also admit that the current practices have some inherent human bias and social expectations, interpretability literature so far has been rich with different useful methods and valuable opinions that we discuss below.
2.2.1 How to Achieve Interpretability?

Interpretability in machine learning models can be achieved in different ways [70]. The first and most preferable approach is to build an inherently interpretable model, e.g., a linear one. However, these models may compromise their predictive capacity for transparent Interpretability. The second approach is to build a model that can perform predictions and simultaneously generate explanations. However, it is a very challenging task because the accepted meaning of the term ‘interpretability’ still needs to be settled in the research community. Moreover, it requires both the ground-truth explanations and the labeled samples to train simultaneously for prediction and explanation generation. The third approach is to use separate explanation methods to work on top of the existing models. That is, the existing models can be any black-box model (e.g., deep learning models), and the explanation methods are responsible for generating explanations for the models. Interpretability is especially important when deep learning models are used for knowledge extraction. Regardless of good predictive performance by a DL model, it may still not be useful for discovery as the model may have only learned spurious correlations [71]. Most of the interpretability methods in the literature are designed around the third interpretability approach, frequently referred to as \textit{post hoc} methods.

2.2.2 Global vs. Local Interpretability

The scope of Interpretability in machine learning is another consideration. For example, \textit{Global Interpretability} deals with the overall behavior of the model, such as discovering
patterns and the interrelationships among them used for predictions. *Global Interpretability* is useful to debug a model, specifically to diagnose if the model has any inherent bias or has learned any artifact instead of the objects of interest. As global Interpretability is very hard to obtain because it requires building a relationship among all predictions made by the model, people traditionally end up with local Interpretability that deals with explaining model behavior case-by-case basis. For example, *Local Interpretability* tries to explain why the image has been classified as "cat"/"dog" or why a particular loan application has been "accepted"/"rejected."

While we recommend reading some other literature reviews [56, 57, 72–76] that cover comprehensive discussion of interpretability methods, we briefly describe the key concepts, axioms, methods and metrics used in interpretable machine learning.

### 2.3 Taxonomy of Model Interpretability Problem

Guidotti et al. [56] divides the black-box explanation into three sub-categories: *model explanation* means explaining the overall logic of the model; *outcome explanation* means finding the correlation between individual input and corresponding decision; *model inspection* means explaining the behavioral change with changes in input and other parameters or explaining what parts of the model take specific micro-decisions. We provide comprehensive insights into the different aspects of the interpretability problem in Figure 2.1.
Figure 2.1: Taxonomy of Explainable AI in terms of transparency and holistic view.

2.4 Important Terminology in Interpretability

As the field of “Explainable AI (XAI)” is growing rapidly, researchers have defined several important notions useful for the discussion. In this section, we discuss several terminologies often considered significant in interpretability literature.

- **Interpretability**: Doshi-Velez and Kim [60] defined interpretability as the “ability to explain or to present in understandable terms to a human." Miler [65] defined this term as “the degree to which an observer can understand the cause of a decision." Gilpin et al. [72] defined interpretability as “...to describe the internals of a system in a way that is understandable to humans...for a system to be interpretable, it must produce descriptions
that are simple enough for a person to understand using a vocabulary that is meaningful to the user." *Interpretability* is a passive characteristic of a model whose only purpose is to make sense to humans, not necessarily clarifying the internals of the model [74].

- **Understandability**: *Understandability* is associated with the notion that if a model’s behavior makes sense to humans without even understanding the mechanistic or algorithmic aspect of the model [77]. *Understandability* is also referred to as *intelligibility* [74].

- **Comprehensibility**: An interpretable model is *comprehensible*, so they imply the same aspect of a model [56].

- **Explainability**: The term *explainability* is associated with the internal mechanisms or learned behavior the model gained during training to make decisions. Gilpin et al. [72] defined *explainability* as “models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions." As pointed out, *explainability* either focuses on the processing of the data or the representation mechanism of the data. The authors argued that explainable models are inherently interpretable, but the reverse is not always true. According to Arrieta et al., *explainability* is an active characteristic of a model [74], denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions.

- **Transparency**: The concept of *transparency* is related to understanding the mechanism by which the model works [61]. According to Lipton, transparency can be at different
levels—at the level of the entire model, the level of the individual components such as input, parameters, and calculation, and the level of the training algorithm.

- **Fidelity:** fidelity of an interpretable model is a comparative assessment of its accuracy with respect to the black-box model the interpretable model is trying to explain [56].

One important point is to note that the terms **interpretability** and **explainability** are elusive. Many studies used the terms interchangeably. However, Gilpin et al. [72] argued that explainability is a more general term, and interpretability is an obvious first step to achieve explainability. However, interpretability alone is not enough and cannot ensure completeness. In particular, the authors emphasized as: “While interpretability is a substantial first step, these mechanisms need to also be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited.”

### 2.5 Axioms of Attribution Methods

Recent interpretability research spelled out some desirable properties of attribution methods as follows:

- **Sensitivity(a):** An attribution method satisfies **Sensitivity(a)** [43] if for every input a and baseline that differ in one feature but have different predictions, then the differing feature should be given a non-zero attribution.

- **Sensitivity(b):** Suppose the function implemented by the deep network does not depend (mathematically) on some variable. In that case, the attribution method is said to be satisfying **Sensitivity(b)** [43] if the attribution to that variable is always zero.
• **Linearity:** Suppose two deep networks modeled by the functions $f_1$ and $f_2$ are linearly composed to form a third network that models the function $a \times f_1 + b \times f_2$, i.e., a linear combination of the two networks. Then we call an attribution method to be satisfying linearity if the attributions for $a \times f_1 + b \times f_2$ to be the weighted sum of the attributions for $f_1$ and $f_2$ with weights $a$ and $b$ respectively [43].

• **Explanation Continuity:** Let $S_c(x)$ be a continuous prediction function for the input $x$ and class $c$. Also, let $x_1$ and $x_2$ be two nearly identical points in the input space, i.e., $x_1 \approx x_2$ for which model responses are identical. Attribution methods, to maintain explanation continuity [77], should generate nearly identical attributions $R^c(x_1)$ and $R^c(x_2)$ i.e., $R^c(x_1) \approx R^c(x_2)$.

• **Implementation Invariance:** Let $m_1$ and $m_2$ be two implementations (models) $S_{m_1}(x)$, $S_{m_2}(x)$ that generate same outputs for the same input $x$: $\forall x : S_{m_1}(x) = S_{m_2}(x)$. An attribution method is called implementation invariant [43] if it generates identical attributions when functions $S_{m_1}(x)$, and $S_{m_2}(x)$ are in the equivalence class for the same input $x$. That is, $\forall (m_1, m_2, x, c.) : R^{c,m_1}(x) = R^{c,m_2}(x)$

• **Sensitivity-n:** An attribution method satisfies sensitivity-$n$ axiom [78] if the replacement of any subset of features by their non-informative baseline causes the output score to drop by the sum of the attributions previously assigned to those features. Let $x_S = \{x_1, x_2, \ldots, x_n\} \subseteq x$ be the subset of features. Then:

$$\sum_{i=1}^{n} R^c(x_i) = S(x) - S(x \setminus x_S) \quad (2.1)$$
This *sensitivity-n* property is only applicable to salience methods that measure the marginal effect of input on the output. Ancona et al. [78] proved that attribution methods (based on gradients), when applied to non-linear models, cannot satisfy *sensitivity-n* property at least for some values of *n*, possibly for the reduced degrees of freedom to capture non-linear interactions.

- **Completeness or Summation to Delta:** This is a variant of the *sensitivity-n*, also called *sensitivity-N*. It constraints the attribution methods to produce attribution that sums equal to the classification score with an assumption that non-informative baseline should produce $S(\bar{x}) \approx 0$. This property is denoted as: $\sum_{i=1}^{N} R^c(x_i) = S(x) - S(\bar{x})$

- **Perturbation - $\epsilon$:** This axiom proposed in [58] is a relaxed version of *sensitivity-1* axiom. Suppose $\{x_1, x_2, \ldots, x_n\}$ be the input features. For a given $0 < \epsilon \leq 1$, if all the features except $x_i$ are fixed, and removal of $x_i$ causes the output to change by $\Delta y$, then the Perturbation - $\epsilon$ is satisfied if the attribution holds the inequality: $\text{attr}(x_i) \geq \epsilon \star \Delta y$.

### 2.6 Taxonomy of Interpretability Methods

In this section, we describe different interpretability methods in the literature. We provide a taxonomy of the interpretability methods in Figure 2.2. We note that this taxonomy is not perfect in the traditional sense, as the categorization of interpretability methods is still evolving. While we discard some infrequent or obsolete approaches and include some emerging methods, this taxonomy is inspired mainly by Ras et al. [57].
1. **Visualization:** Visualization methods focus on highlighting the discriminative regions of the input that mainly influenced the model’s decision. This approach is prevalent for deep learning models, especially in computer vision.

2. **Distillation:** Distillation methods focus on building a separate "transparent box" model, which is directly interpretable to extract the salient regions or crucial decision rules that guide the original model to reach its decisions. Methods under this category are usually model-agnostic. Moreover, the resulting explanations may be a set of rules or visualization of important regions, similar to visualization methods.

3. **Intrinsic:** Intrinsic methods consider model interpretability during model design or training. This approach usually leads toward joint training for predictions and explanations or provides a more transparent model where an explanation is somewhat intuitive. A separate post hoc analysis may be required for the latter ones.

4. **Counterfactual:** Counterfactual explanations [79, 80] usually do not explain the specific output. Instead, it explains in the form of hypothetical scenarios, potentially intending to provide algorithmic recourse. It provides a better understanding of how the decisions change over the input space and allows users more options to change the model’s decision [81].

5. **Influence Functions:** To generate an explanation for a prediction, influence functions [39] find the influence of the training points on the learning algorithm that leads toward this model prediction.
Figure 2.2: Taxonomy of Explainable AI. We also show how frequently these methods have been used in neuroimaging studies.

To precisely define the interpretability methods, we define an input as a vector $\mathbf{x} \in \mathbb{R}^d$. We also define the model as a function $F : \mathbb{R}^d \rightarrow \mathbb{R}^C$, where $C$ is the number of classes in the downstream classification problem. Moreover, let us also assume that the mapping $F_c(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ defines the class-specific logit, where $c$ is the predicted class. An explanation method in computer vision tasks generates an explanation map $E : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that maps
\( x \) to a saliency map of the same shape, highlighting the important regions influencing the prediction.

### 2.7 Visualization Methods

As defined earlier, visualization methods highlight the most influencing regions of the input that drive the model’s output. Generally, visualization methods for model interpretability fall under two main categories. The first category is *Backpropagation Methods*, also called *Sensitivity Methods*, and the latter category is *Perturbation-Based Methods*, also called *Salience Methods* [70]. Though other methods (e.g., LIME and SHAP) may still use visualizations to communicate explanations, we omit them from the visualization category because they require a separate interpretable model to generate explanations.

#### 2.7.1 Gradient-Based Methods

Backpropagation methods are further classified into gradient backpropagation and modified backpropagation methods based on how backpropagation is performed during the computation of saliency maps.

##### 2.7.1.1 Gradient Backpropagation

In gradient backpropagation, also called *sensitivity methods*, we measure how the output score changes with the tiny change in each feature dimension. The sensitivity methods assume this change rate indicates the importance of the corresponding input dimension.
• **Gradients (GRAD):** Gradient (GRAD) [52, 53] is the gradient of the class-specific logit with respect to input features \( x \). Mathematically, \( e = \nabla_x \mathcal{F}_i(x) \), where \( e \) is the vector representing the feature importance estimate for each input variable in the sample. In fact, it determines the input features for which the least perturbation will end up with the most change in the target response. However, gradients are usually noisy indications of attribution [44, 82, 83]. The major pitfall of using gradients is that the partial derivative \( \partial \mathcal{F}_i(x)/\partial x_k \) is not independently related with \( x_k \) but also with other input dimensions. Furthermore, the concept of saliency does not apply to the linear classifier because saliency is independent of the input for linear models.

• **Gradient \( \odot \) Input:** Gradient \( \odot \) Input [84] was introduced to improve the sharpness of the attribution maps obtained through sensitivity analysis. However, Ancona et al. [78] showed that Gradient \( \odot \) Input becomes equivalent to DeepLIFT and \( \epsilon \)-LRP, if the network has only ReLU activation functions and no additive biases. This point-wise multiplication was initially justified to sharpen the gradient explanations. However, it is better justified when the measure of salience is a priority over mere sensitivity [70].

• **Integrated Gradients (IG):** Integrated Gradients [43] is an attribution method that satisfies *implementation invariance* and gives one estimate per feature. IG uses the interpolation technique to integrate importance at different discrete intervals between uninformative baseline, say \( \bar{x} \) and the input \( x \), to give an integrated estimate of feature
importance. The feature importance based on integrated gradients is computed as follows:
\[ e = (x - \bar{x}) \times \sum_{i=1}^{k} \frac{\partial F_i(\bar{x} + \frac{i}{k} \times (x - \bar{x}))}{\partial x} \times \frac{1}{k} \]  
(2.2)

The ultimate estimate \( e \) depends on the value of \( k \) (number of intervals) and the choice of a suitable uninformative baseline \( \bar{x} \). IG also satisfies sensitivity-\( N \) axiom since
\[ \sum_{i=1}^{n} R_c(x_i) = F_i(x) - F_i(\bar{x}) \]

- **Smooth-Grad (SG):** Smoothgrad [44, 85] expresses a feature as an averaging of \( N \) noisy estimates obtained when input is perturbed with some Gaussian noise \( \epsilon \), expressed as:
\[ e = \frac{1}{N} \sum_{j=1}^{N} \nabla_{x+\epsilon} F_i(x + \epsilon), \text{ where } \epsilon \sim N(0, 1) \]  
(2.3)

Other variants [86] of smooth-grad, especially their squared and variance versions, exist in the literature. However, their usage is very limited in model interpretability.

- **CAM and GRAD-CAM:** Zhou et al. [87] proposed *Class Activation Map (CAM)* to visualize the focal regions using global average pooling on the last layer activations in convolutional neural networks. Subsequently, Selvaraju et al. [50] proposed a gradient-weighted class activation map called Grad-CAM and generalized the CAM computation to a broader set of networks by leveraging the gradients of the last layer activation maps. Indeed, Grad-CAM computes the gradients of the class score (logit) with respect to the last convolution layer. Let \( A^k \) be the set of feature maps of size \( m \times n \). Grad-CAM computes
\[ \alpha_k^i = \frac{1}{m \times n} \sum_{i}^{m} \sum_{j}^{n} \frac{\partial y}{\partial A_{i,j}^k}, \]  
the gradients of the output with respect to each feature map, and use average pooling of the gradients to assign a score to the feature map. Finally, it takes
the weighted combination of the feature maps followed by ReLU only, i.e., \( \text{relu}(\sum_k \alpha_k A^k) \),
to consider the positive influence on the class of interest. As Grad-CAM visualization is in the feature map space, Grad-CAM explanation is first upsampled to the input resolution using bilinear interpolation and then overlaid on the input image. Grad-CAM is sometimes combined with Guided backpropagation for pixel-space visualization through an element-wise product called Guided Grad-CAM. Several variants of Grad-CAM, such as GRAD-CAM++ [88] and Score-CAM [89], have been proposed to improve upon Grad-CAM.

Kapishnikov et al. also proposed two approaches [58, 90], called eXplanation with Ranked Area Integrals (XRAI) and Guided IG, that can refine the results of integrated gradients and can produce improved explanations. However, their usage in neuroimaging studies is still minimal.

2.7.1.2 Modified Backpropagation

Modified backpropagation category refers to the methods that use different forms of backpropagation other than standard backpropagation. The modification can be based on how gradients should flow backward when the ReLU layer is encountered, such as in guided backpropagation and DeConvNet methods. Another trend is to use relevance backpropagation instead of gradients, such as in layer-wise relevance propagation and deep Taylor decomposition methods.
• **Guided Backpropagation (GBP):** Guided backpropagation [51] modifies the gradients during backpropagation to make it consistent with ReLU activation functions. Let \( \{f^l, f^{l-1}, \ldots, f^0\} \) be the input and output features maps of the ReLU activations during the forward pass of a DNN. Also, let \( \{R^l, R^{l-1}, \ldots, R^0\} \) be the intermediate gradients during the backward propagation. Precisely, the forward ReLU function at the intersection of \( l-1 \) and \( l \)-th layers is defined as \( f^l = \text{relu}(f^{l-1}) = \max(f^{l-1}, 0) \) and Guided backpropagation overrides the gradients of ReLU functions. The unique purpose of this modification is to allow only non-negative gradients during backpropagation. Mathematically,

\[
R^l = 1_{R^{l+1} > 0}1_{f^l > 0}R^{l+1}
\]  

(2.4)

That is, GBP considers only positive activations with respect to ReLUs and positive gradients from the earlier step during backward propagation.

• **DeConvNet:** DeConvNet [51] is another "guided" method but slightly differs from the Guided Backpropagation in that it only passes "positive" gradients from the upper to the lower layer when the ReLU layer is encountered. The use of DeConvNet to interpret models in neuroimaging domain is very limited.

• **Layer Relevance Propagation (\( \epsilon \)-LRP):** Layer relevance propagation [54] uses the term "relevance" denoted as \( r_i^{(l)} \) to refer to the relevance of the unit \( i \) in layer \( l \). It starts at target neuron \( c \) in the last layer \( L \) and treats the target neuron’s activation as its relevance. The relevance of all other neurons in layer \( L \) are set to 0. Subsequently,
during backward propagation, it computes attributions for neurons at other layers using a recursive \(\epsilon\)-rule as described in Eq. 2.6. Let \(z_{ij} = x_i^{(l)} w_{ij}^{(l,l+1)}\) be the weighted activation of unit \(i\) in layer \(l\) onto neuron \(j\) in the next layer, \(b_j\) be the additive bias for the unit \(j\) and \(\epsilon\) be the small numerical constant to ensure stability. The final attribution for the \(i\)-th input is defined as \(R_i^c(x) = r_i^{(1)}\).

\[
\begin{align*}
  r_i^{(L)} &= \begin{cases} 
  f_i(x) & \text{if unit } i \text{ is the target neuron} \\
  0 & \text{otherwise}
\end{cases} \tag{2.5}
\end{align*}
\]

Layer relevance scores are backpropagated and distributed according to the following rule:

\[
  r_i^{(l)} = \frac{z_{ij}}{\sum_j z_{ij} + b_j + \epsilon \cdot \text{sign}(\sum_j z_{ij} + b_j)}r_j^{(l+1)} \tag{2.6}
\]

Ancona et al. \[78\] showed that \(\epsilon\)-LRP is equivalent to the feature-wise product of the input and the modified partial derivative. Another variant of the original LRP to combat the numerical stability is called \(\beta\)-LRP \[83\].

- **DeepLIFT Rescale**: DeepLIFT (Deep Learning Important FeaTures) assigns attributions to each unit \(i\) based on activations using original input \(x\) and baseline input \(\bar{x}\) \[55\]. Similar to LRP, DeepLIFT Rescale assigns attribution through backward propagation. Let \(\bar{z}_{ij}\) be the weighted activation of neuron \(i\) in layer \(l\) into neuron \(j\) in the next layer and defined as \(\bar{z}_{ij} = \bar{x}_i^{(l)} w_{ij}^{(l,l+1)}\). The rule for assigning attributions during the backward pass is described in Eq. 2.8. The intended attribution for the \(i\)-th input is defined as
\( R^c_i(x) = r^{(1)}_i \). Baseline reference values are created based on a forward pass with input \( \bar{x} \).

\[
\begin{align*}
    r^{(L)}_i &= \begin{cases} 
    \mathcal{F}_i(x) - \mathcal{F}_i(\bar{x}) & \text{if unit } i \text{ is the target neuron} \\
    0 & \text{otherwise}
    \end{cases} 
\end{align*}
\] (2.7)

The attributions are backpropagated according to the following rule:

\[
    r^{(l)}_i = \sum_j \frac{z_{ij} - \bar{z}_{ij}}{\sum_{i'} z_{i'j} - \sum_{i'} \bar{z}_{i'j}} r^{(l+1)}_j 
\] (2.8)

DeepLIFT Rescale generalizes the concept of \( \epsilon \)-LRP with no assumption of the baseline or a particular choice of non-linearity. In other words, \( \epsilon \)-LRP becomes equivalent to DeepLIFT if the baseline is 0 and only ReLU or Tanh is used in the network with no additive biases. DeepLIFT and \( \epsilon \)-LRP replace the gradient of the non-linearities with their average gradient. However, this replacement does not apply to discrete gradients. Hence the overall computed gradient of the function may not be the average gradient of the function as a whole. Due to this constraint, DeepLIFT and \( \epsilon \)-LRP do not satisfy implementation invariance. DeepLIFT was originally designed for feed-forward networks, and Ancona et al. [78] showed that DeepLIFT is a good approximation of Integrated Gradients for feed-forward networks.

- **Deep Taylor Decomposition:** Montavon et al. [82] proposed another relevance backpropagation approach to pass relevance from the output to the input space. This backpropagation of relevance is similar to LRP but uses a different formulation using first-order Taylor expansion.
2.7.2 Perturbation-Based Methods:

In perturbation-based methods, also called salience methods, the marginal effect of a feature on the output score is computed relative to the same input where such a feature is absent.

- **Occlusion Sensitivity:** Zeiler and Fergus (2014) [48] proposed a perturbation-based approach called *Occlusion Sensitivity* to measure the sensitivity of the output score when some regions in the input image are occluded. This approach is also known as *Box Occlusion* because of using a grid or box structure during occlusion. Precisely, this method occludes different portions of the input with a grey square and expects a significant drop in classification score if the portion is strongly discriminative for the prediction the model has made.

- **Meaningful Perturbation:** Fong and Vedaldi (2017) [42] proposed a model-agnostic generalization of gradient-based saliency that uses input perturbations and integrates information obtained through all backpropagation. Suppose the input image be $x_0$ and $f(x) \in \mathbb{R}^C$. The goal is to find the smallest deletion mask $m : \Lambda \rightarrow [0, 1]$ for which the classification score drops very significantly, i.e., $f_c(\Phi(x_0; m)) \ll f_c(x_0)$, where $\Phi(x_0; m)$ is perturbation operator. The problem of finding the minimum deletion mask is defined as the following optimization problem:

$$
m^* = \arg \min_{m \in [0,1]^\Lambda} \lambda \|1 - m\|_1 + f_c(\Phi(x_0; m)) \quad (2.9)$$

$\lambda$ is a regularizing parameter that enforces small deletion to generate a highly informative region to explain the prediction. This optimization problem is solved using the gradient
descent technique.

Gradient-based methods are fast, easy to implement, and readily applicable [43] to existing models compared to perturbation-based methods. However, gradient-based methods are extremely noisy, usually affected by high-frequency variations, and may not represent the model’s decision-making process. In contrast, perturbation-based methods are directly interpretable (because it computes the marginal effect), model-agnostic, and do not require accessing the internal operations of the models.

While the major advantage of perturbation-based methods is the direct computation of the marginal effect of each feature or a small subset of features, the obvious limitations are that the perturbation methods are very slow compared to gradient-based methods. Moreover, they must choose the number of input features to perturb at each iteration and the perturbation technique because the explanations depend heavily on these hyperparameters. Ideally, for realistic reasons, it is not possible to test perturbations of all possible subsets. Moreover, there is no rigorous theoretical foundation to choose from the available perturbation techniques, thus making the explanations unreliable.

2.8 Distillation Methods

In distillation methods, a separate explanation model, also called interpretable model, is required to explain the decision of the original model. This approach is model-agnostic, and the interpretable model does not need the internal behavior of the model. As a separate model is used to extract the essential aspects of the original model, this process is called
distillation. However, similar to visualization methods, distillation methods may still produce visualization as explanations.

- **LIME**: LIME [49], also called *Local Interpretable Model-agnostic Explanations*, is based on a surrogate model. The surrogate model is usually a linear model constructed based on different samples of the main model. It does this by sampling points around an example and evaluating models at these points. LIME generally computes attribution per sample basis. It takes a sample, perturbs multiple times based on random binary vectors, and computes output scores in the original model. It then uses the binary features (binary vectors) to train an interpretable surrogate model to produce the same outputs. Each of the coefficients in the trained surrogate linear model serves as the input feature’s attribution in the input sample. Let $x = h_x(x')$ be a mapping function between "interpretable inputs" ($x'$) and "original inputs" ($x$). Also, let $x' \in \{0, 1\}^M$, $M$ be the number of simplified features, and $\phi_i \in \mathbb{R}$. The local interpretable explanation model is defined as:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i$$  \hspace{1cm} \text{(2.10)}$$

The explanation model $g$ can be obtained by solving the following optimization problem:

$$\xi = \arg \min_{g \in \mathcal{G}} L(f, g, \pi_{x'}) + \Omega(g)$$  \hspace{1cm} \text{(2.11)}$$

$g(x')$ and $f(h_x(x'))$ are enforced to be equal. That is, $L(f, g, \pi_{x'})$ determines how unfaithful $g$ is when it approximates $f$ in the vicinity defined by the similarity kernel $\pi_{x'}$. $\Omega$ penalizes the complexity of $g$ and the Equation 2.11 can be solved using penalized
linear regression. One of the major issues with LIME is robustness. LIME explanations can disagree if computed multiple times. This disagreement occurs mainly because this interpretation method is estimated with data, causing uncertainty. Moreover, the explanations can be drastically different based on kernel width and feature grouping policies.

• **SHAP**: Historically, Shapley values are computed in a cooperative game theory to calculate the marginal contributions of each player. The computation of this marginal effect relies on game outcomes of all possible sets of coalitions. Suppose $P$ be a set of $N$ players and a function $\hat{f}$ that maps any subset $S \subseteq P$ of players to a game score $\hat{f}(S)$. This score is obtained when the subset $S$ of players participated in the game. The Shapley value is a way to compute the marginal contribution of each player $i$ for the game outcome $\hat{f}(P)$—the outcome obtained when all players $P$ participated in the game.

$$R_i = \sum_{S \subseteq P \setminus \{i\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!} [\hat{f}(S \cup \{i\}) - \hat{f}(S)]$$  \hspace{1cm} (2.12)

The problem with Shapley values is that this attribution technique is computationally intractable when the number of players is large. Lundberg and Lee [47] proposed a regression-based, model-agnostic formulation of Shapley values called SHapley Additive exPlanations (SHAP). This approach is also known as Kernel SHAP and is widely used to compute SHAP explanations. As SHAP ranks the features based on their influence on the prediction function, the occurrence of overfitting is usually reflected in the provided explanation. In fact, Kernel SHAP removes the need to use heuristically chosen parameters
as used in LIME to recover SHAP values.

LIME and SHAP could also be treated as perturbation-based methods because they both perturb the original input locally to build separate interpretable models. However, as described here, the category of perturbation-based methods does not rely on a separate interpretable model. Hence, LIME and SHAP belong to a separate category for their model-agnosticism and the usage of a separate model.

2.9 Intrinsic Methods

Intrinsic methods focus on interpretation as part of the model design or training rather than doing a separate post hoc analysis. These methods are model-specific and are usually implemented based on different design or training perspectives. While some shallow models, such as linear models and decision trees, are directly interpretable, deep learning models are considered black boxes, and their internal functions are quite inscrutable. However, there are many doors to obtain intrinsic interpretability in DL, such as attention mechanism, joint training, and modular transparency. In this section, we briefly discuss some of the common practices used to obtain intrinsic interpretability in deep learning.

2.9.1 Attention Mechanism

An attention mechanism is a technique generally used in deep learning models which computes the conditional distribution over inputs leading to a vector of weights that specify the importance of different regions in the input for the given context. There are several approaches [91, 92] to compute attention weights for single-modal or multi-modal tasks. The
attention mechanism has been proven to improve the deep learning model’s performance, and attention weights can be visualized as heatmaps to provide easy-to-understand explanations.

2.9.2 Joint Training

*Joint training* is the concept of training a model simultaneously for performance and explanations [93–95]. Joint training requires a complex objective function to optimize for the additional explanation task. The additional task may provide a direct textual explanation, generate an explanation association between inputs or latent features and human-understandable concepts, or learn semantically meaningful model prototypes [96]. A very high-level view of joint training optimization can be as follows:

$$\arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \alpha L(y_n, y'_n) + L(e_n, e'_n)$$  \hspace{1cm} (2.13)

The arguments $y_n$ and $y'_n$ refer to model output and output label, respectively. $e_n$ and $e'_n$ refer to model explanation and explanation label, respectively.

2.9.3 Modular Transparency

*Modular transparency* [97] refers to a network consisting of multiple modules. The modules have pre-specified design goals and are usually black-boxes. However, the interaction among the modules is transparent. The explanation can be obtained from understanding how the model functions globally. Ba et al. [98] demonstrated a modular deep learning model constructed with attention mechanism and reinforcement learning for multiple object recognition tasks. The model was inspired by how humans perform visual sequence recognition tasks.
by continually moving to the next relevant locations, recognizing individual objects, and changing the internal sequence presentation.

2.10 Counterfactual Explanations

Counterfactual explanations, by definition, provide explanations for hypothetical scenarios. Specifically, counterfactual explanations simply ask for the smallest change required to change the model’s outcome. This category of explanations is human-friendly [99] because they allow humans to choose from multiple options to change the scenarios. Wachter et al. [79] proposed a single-objective optimization method to generate a counterfactual explanation.

\[ L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x') \]  

(2.14)

The inequality \(|\hat{f}(x') - y'| \leq \epsilon\) determines the tolerance between the current and the counterfactual predictions. The parameter \(\lambda\) balances the distance in prediction and the distance between original and counterfactual instances. \(d(x, x')\) is the distance between the original instance \(x\) and the counterfactual \(x'\) measured as weighted Manhattan distance as defined below:

\[ d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j} \]  

(2.15)

where \(MAD_j\) is the median absolute deviation of feature \(j\). Dandl et al. [81] proposed a multi-objective formulation of counterfactual explanations. This multi-objective formulation satisfies multiple requirements of counterfactual explanations. Other implementations of
counterfactual explanations can be found in [80, 100].

2.11 Influence Functions

Studies also proposed a data modeling approach to explaining a model prediction in terms of influence functions [39, 101]. Precisely, these methods attempt to find the representative training samples that influenced the prediction of the test sample. While this area of investigation toward explainability is still at the rudimentary level, few studies [39, 101–108] proposed approaches to determine the influencing training points for a particular test case. While determining influence function is yet to use in neuroimaging research as far as we know, this approach, if carefully leveraged, can lead toward many advantageous use cases, including generating counterfactual explanations [101] for different neurological disorders.

2.12 Sanity Checks for Interpretability Methods

It is generally expected that model explanation methods should be reasonably sensitive to model parameters. Moreover, the people expect that model should map data and the associated label based on the data generation mechanism relevant to the target. So, to understand if the behavior of an explanation method is reasonable or not, Adebayo et al. [109] proposed the following sanity checks:

**Model Randomization Test:** As a model goes through an intensive training process and learns its parameters during the training process, explanations must be sensitive to the model parameters. For this kind of model randomization test, people use either full randomization or cascading randomization and expect to have varied explanations from the explanations
generated using the original (non-randomized) model.

**Data Randomization Test:** In this test, training labels are permuted to break the relationship between data and associated labels. A model is trained on these shuffled data and forced to memorize the labels against each training sample. As the model memorizes rather than learning the inherent logical, structural, or causal relationship between data and labels, it performs no better than a random model during inference. However, for any plausible explanation method, the post hoc explanation of this model should be substantially different from the model trained on the original training data. However, this test is extremely time-consuming because a model trained on randomized data takes a long time and customized hyperparameters to achieve reasonable convergence.

### 2.13 Evaluation Metrics

Human evaluation (qualitative) of explanation methods can be entirely wrong because it is possible to create adversarial samples \[^{[110, 111]}\] that can fool the human eye, totally changing the model predictions. For quantitative assessment, we need to define the domain-specific desired properties of the interpretability methods formally. Moreover, we need appropriate quantitative metrics to assess the behavior of an interpretability method. When the generated attributions do not become plausible, it is hard to identify if the problem is due to the model itself or to the interpretability method that generated the attributions. In this section, we present some evaluation metrics proposed in the interpretability literature.
2.13.1 Metrics for Ground-truth Datasets

Arras et al. [112] proposed two evaluation metrics that can reliably quantify the explanation methods for the datasets that have ground truths.

**Relevance Mass Accuracy:** This metric calculates the proportions of total attributions that reside within the relevance area.

\[
\text{Relevance Mass Accuracy} = \frac{R_{\text{within}}}{R_{\text{total}}}
\]

where \( R_{\text{within}} = \sum_{k=1}^{\vert GT \vert} r_{p_k} \) and \( R_{\text{total}} = \sum_{k=1}^{N} r_{p_k} \) (2.16)

where \( r_{p_k} \) is the relevance score for the pixel \( p_k \). \( N \) is the total number of pixels. \( GT \) is the set of all pixels within the relevance area (ground-truth area).

**Relevance Rank Accuracy:** Let \( K \) be the number of pixels within the ground truth masks. This metric measures how many high-ranked \( K \) pixels are within the relevance area. Let \( P_{\text{top } K} = \{ p_1, p_2, \ldots, p_K \mid r_{p_1} > r_{p_2} > r_{p_3} \cdots > r_{p_K} \} \) be the top \( K \) pixels sorted in descending order of their attribution values. \( \text{Rank Accuracy} \) is defined as follows:

\[
\text{Relevance Rank Accuracy} = \frac{\vert P_{\text{top } K} \cap GT \vert}{\vert GT \vert}
\] (2.17)

The argument \( GT \) refers to the set of pixels within the ground-truth region.

2.13.2 Metrics for Real Datasets

Several studies proposed different measures, such as Remove And Retrain (ROAR) [86], RemOve And Debias (ROAD), Accuracy Information Curves, Softmax Information Curves [58], Infidelity, Sensitivity [113], to assess the quality of explanations.
**Remove and Retrain (ROAR):** Hooker et al. [86] proposed another approach to evaluate the performance of an interpretability method. In this approach, samples are modified based on the post hoc explanations. In particular, the features that receive significant attributions during explanation are removed. The model is trained over the modified training data, and people expect a sharp drop in model performance because important discriminative features are absent from the training data. The method is time-consuming as it requires full retraining of the model. Another pitfall of this evaluation process is that the ROAR metric may produce erroneous evaluations when correlations among features exist and capturing only the subset of correlated features is sufficient for correct prediction [114]. However, ROAR fails to evaluate the feature relevance correctly in that scenario.

**log-odds score:** Shrikumar et al. [55] proposed a metric to evaluate the quality of explanations. This method greedily identifies the main contributing pixels to convert the original prediction \(c_0\) to some target prediction \(c_t\). That is, it removes pixels (20% of the image) based on descending ranking of \(S_{c_0} - S_{c_t}\). Finally, it measures the change in the log-odds score between \(c_0\) and \(c_t\) for the original image and the image with pixels removed to get the prediction \(c_t\). The greater change in log-odds score implies the greater significance of the removed pixels for the original class and thus better capture the true importance. This metric is not useful for natural images and possibly meaningful for images with a strong structural association as in MNIST.

**Area Over MoRF Precision Curve:** Samek et al. [83] proposed an evaluation technique for the heatmaps based on the idea of how quickly the function value \(f(x)\)
(probability score) drops if the most relevant regions are perturbed. To achieve this agenda, it creates an ordered set $0 = (r_1, r_2, \ldots, r_L)$ based on the importance scores of pixels as assigned by the interpretability method. This procedure follows a region perturbation (most relevant first (MoRF)) process, where gradually, a small rectangular region $m \times m$ surrounding each important pixel location $r_p$ is removed by the uniform distribution. The quantity of interest here is termed as Area Over MoRF Perturbation Curve (AOPC).

$$\text{AOPC} = \frac{1}{L + 1} \left( \sum_{k=0}^{L} f(x_{\text{MoRF}}^{(0)}) - f(x_{\text{MoRF}}^{(k)}) \right)_{p(x)}$$

Here \( \langle \cdot \rangle_{p(x)} \) indicates average over all samples in the dataset. The intuition is that if the ranking strongly associates with the class label, the removal will cause a steeper drop in the functional value, causing a larger AOPC.

Though localization and saliency have different connotations, the quality of a saliency map is often measured as its localization accuracy because they overlap. For example, for a dog image, the localization box usually encapsulates the entire dog without focusing on salient details of the dog. The usage of localization in saliency evaluation is often referred to as weakly supervised localization because neither model training nor post hoc interpretability use localization information.

**Smallest Sufficient Regions (SSR):** Dabkowski et al. [115] proposed a metric based on the notion of the smallest sufficient region capable of correct prediction. This metric requires maintaining the same classification and finding the smallest possible area of the image. This metric is formally defined as follows:
\[
s(a, p) = \log(\tilde{a}) - \log(p)
\]  \hspace{1cm} (2.18)

\[\tilde{a} = \max(a, 0.05),\] where \(a\) is the proportion of the cropped image to the original image. \(p\) is the probability of the corresponding object class when the classifier classifies based on the cropped but resized image. The lower value of \(s(a, p)\) indicates a better saliency detector because it directly translates the idea of SSR — less area, greater probability score. However, this metric is not suitable if the model is susceptible to the scale and aspect ratio of the object. Moreover, as this metric depends on rectangular cropping and reports results as a function of the cropped area, this approach highly penalizes if the saliency map is coherently sparse [58]. Because, in that case, it may span a larger area of the image than the map, which is locally dense, even with the same number of pixels. However, this is counterintuitive from the human vantage point. Humans tend to have sparse and coherent explanations. Moreover, this imposes a severe challenge because masking creates a sharp boundary between the masked and salient region, causing an out-of-distribution problem for the model.

**RemOve And Debias (ROAD):** Rong et al. [116] proposed an evaluation strategy that overcomes the 99% computational cost of retraining to evaluate attribution methods. The authors made a useful experimental observation that existing ROAR evaluations based on MoRF (most relevant first) or LeRF (least relevant first) removal strategies are inconsistent in ranking the attribution methods. The authors attributed this inconsistency to the class information leakage through the shape of the removed pixels. To mitigate these unwanted influences, the authors proposed a *Noisy Linear Imputation* operator that debiases the
masking effect and removes the need for additional retraining.

Recently, Kapishnikov et al. proposed another perturbation-based evaluation metric [58], called *Performance Information Curve* to evaluate the appropriateness of an attribution method.

**Performance Information Curve (PIC):** The PIC evaluation builds a saliency-focused image. It starts with a blurred image and combines with a saliency mask thresholded, for example, at x%, to produce the saliency-focused image. The saliency-focused image is then fed into the model to assess the performance of the attribution. The accuracy/softmax score of the model is then mapped as a function of *Information Level*, i.e., calculated entropy. The entropy is a proxy measure of the information content re-introduced for evaluation. The compressed image size is an approximate proxy for the information content of an image. It normalizes the entropy of the re-introduced image by considering the proportion of the entropy from the original image. The aggregate performance measurement over all the information levels for all samples in the dataset finally generates the PIC. The PIC has two variants:

**Accuracy Information Curve (AIC):** For AIC, the x-axis uses normalized entropy values and divides them into several bins. The y-axis reports the accuracy calculated over all the saliency-focused images for each bin of image information level (entropy).

**Softmax Information Curve (SIC):** The x-axis uses the same normalized entropy values for SIC. The y-axis reports median scores for the proportion of the original label’s softmax score for the saliency-focused image versus the softmax for the original image.
2.14 Criticisms of Post hoc Interpretability

The concept of interpretability is simultaneously considered essential and evasive \([61, 117]\). A vast amount of studies \([41, 118–131]\) talked about different pitfalls of post hoc interpretability methods. Those studies have diverse opinions about the transparency of deep learning models and the applicability of popular interpretability methods in real-world deployment scenarios. For example, Rudin (2019) \([41]\) criticized attempts to explain black-box models. Instead, she suggested building inherently interpretable models. Rudin also thinks black-box models are not required in AI \([120]\). Studies also have focused on different aspects of those interpretability methods to determine their reliability and efficacy. For example, multiple studies \([40, 41, 132]\) have demonstrated that saliency maps are unreliable for localizing abnormalities in medical images.

While post hoc interpretability methods have been widely used in different applications and neuroimaging studies, we should always be aware of their usage when safety and trust are our significant concerns. For example, we must accept the explanations wisely when we want to use interpretable DL models to understand how the brain functions or what dynamics are responsible for a particular mental disorder. Generally, people use post hoc interpretability methods without any pre-condition applied to the model’s design. Paez \([131]\) argued that model transparency or model approximation is useful for objectively understanding the model. Moreover, it is also a necessary condition to achieve post hoc interpretability.
2.15 Conclusion

In this chapter, we introduce the problem of interpretability or explainability for AI models from a holistic point of view. We discuss the desiderata of interpretability in AI, the philosophical views of scientific explanations, and the axioms that need to be satisfied by the interpretability methods. We provide a helpful taxonomy of interpretability methods and indicate their usage trends in neuroimaging research. We also discuss different methods, the sanity tests to justify their initial applicability, and the evaluation metrics useful to run interpretability experiments on synthetic and real datasets.
CHAPTER 3
DEEP LEARNING INTERPRETABILITY IN NEUROIMAGING

Deep learning models have been popular due to their ability to learn directly from the raw data in an end-to-end paradigm, alleviating the concern of a separate error-prone feature extraction phase [6]. Nevertheless, the challenges of deep learning models still exist because transparency in these models still needs to be improved to deploy these models in safety-critical domains such as healthcare. In recent years, Explainable AI (XAI) has undergone a surge of developments [57, 133] mainly to get intuitions of what the models have learned and how the models reached the decisions. While the interpretability domain is advancing rapidly, we still need rigorous methods and validation techniques to deploy these models effectively. This chapter comprehensively discusses interpretable deep learning models in the neuroimaging domain and presents our analyses and suggestions for future practices.

3.1 Introduction

Psychiatric disorders have strong correspondence with underlying complex brain dynamics. These ever-changing dynamics supposedly reflect the progression of these disorders. Identifying the essential, interpretable, non-invasive imaging biomarkers from the dynamics can be a significant breakthrough for early diagnosis, potentially preventing its future progression with the help of new insights the model can gain from the data. However, traditional machine learning algorithms mostly rely on hand-crafted features and cannot learn from high-dimensional data, resulting in a drastic performance drop in predictive tasks. Extracting hand-crafted features usually requires the involvement of domain experts in the learning
pipeline and relies on unwarranted assumptions.

One of the crucial challenges of Neuroimaging research is understanding the association between cognitive state and the underlying brain activity [18, 134]. Traditionally, people use the feature engineering approach with shallow linear interpretable models to tackle these challenges.

3.1.1 Feature Engineering Approach to Neuroimaging

In this section, we discuss the traditional feature engineering and feature learning practices in neuroimaging studies. Feature engineering or feature selection step intends to reduce the dimension of the signals while preserving useful discriminative information. Global feature-based (voxel-based) or regional feature-based approaches are commonly used in neuroimaging for feature selection [22]. Ashburner and Friston [7] summarized the advances of voxel-based morphometry (VBM), where voxel-wise parametric statistical tests are conducted to compare the smoothed gray-matter images from the two groups. Kloppel et al. [135] used normalized grey matter segment to classify AD patients from normal cohorts. Saima et al. [136] used the volume of gray matter (GM), the volume of white matter (WM), the volume of cerebrospinal fluid (CSF), the area of the left hippocampus, and the area of the right hippocampus to classify AD from sMRI images based on an ensemble of classifiers. Schnack et al. [137] used gray matter densities (GMD) to model SVM for schizophrenia and bipolar classification using sMRI images. Patel et al. [20] proposed a stacked autoencoder for schizophrenia classification. The autoencoder was trained in an unsupervised fashion on 116 active gray matter regions to extract region-specific features. Subsequently, the extracted features were
used to train an SVM model. Dluhovs et al. [12] used three imaging features (gray matter, white matter, and modulated GM and WM tissue segments of sMRI scans to feed into SVM classifiers in a distributed setting. Xiao et al. [138] used the cortical thickness and surface area features of 68 cortical regions from sMRI images for the SVM-based classification of schizophrenia. Steele et al. [139] used mean grey matter volume and density across 13 paralimbic regions of sMRI scans in SVM based classifier to predict psychopathic traits in adolescent offenders. The regional feature-based approaches intend to summarize the whole brain signal by extracting features from some predetermined regions of interest (ROIs). For example, several studies [140, 141] divided the whole brain into multiple regions and extracted features from those regions to train machine learning models. The ROIs are predetermined based on prior neurobiological knowledge relevant to the disorders.

Rashid et al. [142] used dynamic brain connectivity from resting state fMRI for schizophrenia and bipolar patients classification and showed that dynamic FNC outperforms static FNC. Iddi et al. [143] proposed a two-stage approach for predicting AD progression. In the first stage, the authors used the joint mixed-effect model for multiple modalities such as cognitive and functional assessments, brain imaging, and biofluid assays with fixed effects for covariates like age, sex, and genetic risk. In the second stage of prediction, a random forest algorithm is used to categorize the panel of predicted continuous markers into a diagnosis of controls and stages of progression. Many other studies [144–146] used functional network connectivity measured as Pearson’s correlation coefficients as features for a range of classifiers. Shen et al. [144] also used locally linear embedding (LLE) to reduce the dimensionality of the feature
space to demonstrate that PCA in place of LLE hardly provides separable data points. For a detailed review of feature reduction techniques, refer to [6].

3.1.2 Deep Learning Approach to Neuroimaging

Feature engineering and shallow models suffer from several limitations: 1) the inherent interpretability of shallow models compromises the capacity to deal with high-dimensional neuroimaging data 2) it prevents the natural understanding of brain dynamics. While standard machine learning models can perform reasonably well on handcrafted features, their performance dramatically drops when trained on raw data because of their inability to learn adaptive features from the raw data [27].

In contrast, Deep Learning (DL) has gained significant progress in different application areas, especially for computer vision and natural language processing tasks. The primary benefit of DL is that it can independently learn from the data through varying levels of abstraction using a series of nonlinear functions. Importantly, it relieves the need to use feature engineering, which predominantly relies on some preoccupations with the data that may prevent the natural emergence of significant features. To leverage the capacity of DL in neuroimaging research, researchers have started using DL to reach a new level of understanding of the association between psychiatric disorders and brain dynamics [16, 25, 30, 32, 134, 147, 148].

However, the improved performance of DL comes at the cost of intelligibility—its decision-making process is quite incomprehensible to human beings. While deep learning methods can simultaneously achieve unprecedented predictive performance and potentially lead to
identifying idiosyncratic brain regions associated with the disorders, the model may overfit and not generalize well to unseen subjects. Moreover, it may learn unexpected artefactual associations for its predictions. The need for explanations arises from inadequate knowledge of the data and associated data generation mechanism and poor understanding of the model’s behavior during training. This lack of intelligibility prevents the widespread deployment of DL models in safety-critical domains such as healthcare, medicine, neuroscience, and self-driving cars, to name a few.

Evidence from many recent studies reinforces the potential of deep learning toward new knowledge discovery in different domains. For example, several studies [37, 38] have demonstrated that a convolutional deep learning model, when introspected with gradients, smoothgrad, and GradCAM, might reveal crucial medical information from ECG signals. Often, interpretability may assist in identifying if the model has inherited any inherent bias from the data. For example, Young, Booth, Simpson, Dutton, and Shrapnel [149] used GradCAM and Kernel SHAP to show that produced saliency maps pass some sanity checks and can be helpful at least to diagnose potential biases in the models trained for melanoma detection. In another study, Vellido [150] pointed out the significance of interpretability and visualization in medicine and healthcare. Lucieri et al. [151] used a concept activation vector (CAV) to show that the deep learning model can encode understandable human concepts and apply the disease-relevant concepts for its predictions in a cancer classification task.

From the perspective of neuroimaging applications, we must meet the two most crucial challenges to gain a broader level of acceptance of DL as a research and clinically supportive
tool: 1) Neuroimaging data is inherently high-dimensional. Studies usually have a small sample size posing $m \geq n$ problem, which is very susceptible to cause overfitting in deep models. 2) DL models are considered as \textit{black box models} because of their multi-level non-linearity and lack of established theory behind their learning mechanism. Consequently, it is hard to establish an association between the predictive cognitive state and the underlying dynamics. In other words, the accuracy may not be representative of the quality of the features used by a model. For example, Lapuschkin et al. [152] demonstrated how a \textit{Fisher Vector} model can learn to choose unintended artifacts for generating predictions. In this specific example, the model used \textit{copyright tag} to predict "horse" as all the horse images contain the copyright tag, which turned out to be a characteristic of horses. This kind of phenomenon is entirely unexpected and must be avoided while leveraging deep learning models in medical domains.

\subsection*{3.2 Related Work}

There exist some reviews in the literature [8, 24, 97, 153] for interpretable deep learning in neuroimaging and medical domains [154, 155]. However, they are either focused on machine learning models or general medical imaging, and very few focus on deep learning interpretability in connection to neuroimaging. Moreover, it needs to be clarified the usage trend of these methods and their utility in clinical practices and scientific discovery. That is, how frequently the most prevailing methods have been utilized in earlier research needs to be clarified. Very little research [156] discussed the desiderata of interpretability framework
in neuroimaging. So, there still remains a scope to make a comprehensive accumulation of the prevailing concepts focusing on aspects of deep learning performance, novel findings in interpretability research, and possible implications and connections between them in neuroimaging domain. This review aims to provide a field guide for interpretable deep learning for neuroimaging study, especially for new aspirants in this direction of research.

3.3 A Quick Reference to Interpretability in Neuroimaging

In this section, we first provide a quick glimpse of the deep learning field and a few instances of interpretability studies in neuroimaging. Next, we provide a detailed review of the contexts to which interpretability was applied and the resulting findings.

"Explainable AI"—a subfield of AI has been very popular because of the recent surge in AI models and algorithms as reflected in the left panel of Figure 3.1. Moreover, deep learning shares a larger part of most recent AI practices. Neuroimaging community has also witnessed a similar surge in deep learning practices in recent years. As DL models are black boxes, the need to interpret the DL models has become essential to validate the models or to advance our understanding of the problem domain, as we can see in the right panel of Figure 3.1. For a quick reference to some neuroimaging studies using popular interpretability methods, we also provide some representative neuroimaging studies in Table 3.1.
Figure 3.1: **Left:** "Explainable AI" is getting popular or becoming an area of concern over the years (2012 - 2022) as reflected in the Google Trends Popularity Index (Max. value is 100). **Right:** To get relevant statistics, we searched with the keywords "deep learning in neuroimaging" and "interpretability in deep learning" at this website [https://app.dimensions.ai/discover/publication](https://app.dimensions.ai/discover/publication) (Accessed on October 13, 2022). Neuroimaging studies increasingly used deep learning models during the last decade (2012 - 2021) to understand the dynamics of brain functions and anatomical structures. The need to interpret black-box models is growing accordingly.
Table 3.1: Illustrative neuroimaging studies using popular interpretability methods

<table>
<thead>
<tr>
<th>Interpretability Method</th>
<th>Visualization</th>
<th>Category</th>
<th>Influence Functions</th>
<th>Counterfactual [79, 81]</th>
<th>Distillation</th>
<th>Intrinsic</th>
<th>Standard Backprop</th>
<th>Modified Backprop</th>
<th>Perturbation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occlusion Sensitivity [48]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Meaningful Perturbation [42]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gradients [52, 53]</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Integrated Gradients [43]</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gradient ⊙ Input [84]</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Grad-CAM [50]</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SmoothGrad [44]</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>DeepLIFT [55]</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>DeConvNet [51]</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Guided Backpropagation [51]</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Deep Taylor Decomposition [82]</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Layer-wise Relevance Propagation (LRP) [54]</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>LIME [49]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SHAP [47]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Attention [91, 92]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Joint Training [93–95]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Model Transparency [98]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

3.4 Transfer Learning in Neuroimaging

Leonardsen et al. [215] proposed a CNN model for brain age prediction and subsequently showed evidence of how a model trained to predict age can learn abstractions of the brain and hence can be useful for a series of downstream tasks. The model was selected from some architectural variants and performed well for brain age prediction. The representations as learned by the model were noticeably predictive compared to a baseline model for different unseen datasets for multiple case-control studies. The authors further studied the deviation of the predicted age from the chronicle age by correlating the brain age delta and different standard measures of MRI images.

Eitel et al. [216] emphasized the significance of transfer learning by showing how learned knowledge can be transferred across diseases (AD to MS) and MRI sequences (MPRAGE to FLAIR). However, we argue that transferring knowledge across diseases can be misleading. That is, transferring knowledge from a model trained on Alzheimer’s patients to a study to classify MS patients may confuse the downstream model. Instead, we should define a pretext task and apply unsupervised or self-supervised pretraining of the model on a more neutral group (e.g., healthy controls). This knowledge transfer approach, as we think, may result in more interpretable knowledge transfer [18].

Rahman et al. [18] proposed a transfer learning mechanism that uses contrastive learning to pretrain a deep learning model on publicly available healthy subjects of the Human Connectome Project (HCP). The authors showed that the self-supervised pretraining improved performance of three downstream models separately trained to classify (schizophrenia,
Alzheimer’s disease, and autism spectrum disorder) patients of three disorders with the diverse demographic background. In addition, the improved representations improved the post hoc interpretability of the models.

Oh, et al. [27] argued in favor of deep learning-based approaches compared to the traditional way of building classical machine learning models based only on feature extraction approaches. They incorporated a transfer learning mechanism to transfer knowledge (weights) learned during AD vs. NC classification for the pMCI (progressive mild cognitive impairment) vs. sMCI (stable mild cognitive impairment) classification task.

For a more detailed review of how transfer learning has been used in magnetic resonance imaging, we refer to the paper [35].

3.5 Review of Interpretability Methods in Neuroimaging

For the comprehensive review, we group the papers based on the interpretability methods used in those studies. As some studies used several methods in a single study, we mention them at all relevant places. The summary of the review can be accessed from Table 3.2.

3.5.1 Gradient-based Methods

CAM/Grad-CAM/Guided Grad-CAM

Yang et al. [176] proposed three approaches for generating explanations. One of them, SA-3DUCM (sensitivity analysis by 3D ultrametric contour map), deals with sensitivity analysis of 3D-CNN via a hierarchical image segmentation approach, and the other two methods (3D-CAM, 3D-GRAD-CAM) generate explanations via visualization of network
activations on a spatial map. The methods have their own constraints and complement each other. As a baseline method, the authors used occlusion using a cubic neighborhood of $7 \times 7 \times 7$. However, these occlusion methods are not semantically meaningful. The neighborhood size is a hyperparameter and can drastically change the results. Moreover, this method is computationally very expensive. To address these issues, the authors used 3DUCM to produce semantically meaningful, hierarchical, and compact brain segments. Subsequently, they used the occlusion technique based on these segments rather than individual voxels. However, this addition to the baseline occlusion does not consider correlations and interaction among segments. To resolve this, they used 3D Class Activation Mapping (3D-CAM) and 3D-Grad-CAM, which still suffer from the low-resolution problem and may miss the fine details of importance score in the input space. Further analysis of heatmaps reveals that occlusion generated heatmaps fail to identify discriminative regions. SA-3DUCM and 3D-CAM are able to identify some regions that match with human expert evaluation.

Hu et al. [217] proposed an interpretable DL framework to classify subjects’ cognitive ability (low/high WRAT groups) from n-back fMRI data from the PNC cohort. The proposed model can learn from multimodal fusion data and preserve the association across modalities. The authors leveraged Grad-CAM to guide convolutional collaborative learning. This study takes advantage of multimodal fusion from brain FC data and single nucleotide polymorphism (SNP) data. This study intends to extract potentially useful brain mechanisms within and between brain FC and genetics. 264 ROIs were used for brain FC data. The genetic SNP data were collected from the Illumina HumanHap 610 array, the Illumina HumanHap 500
array, and the Illumina Human Omni Express array.

The results show that the classifier based on convolutional collaborative learning outperforms the traditional ML classifiers. While it has been evident that all classifiers used some hand-engineered features, the low performance of traditional classifiers might arise from the dimensionality reduction of the original hand-engineered Brain FCs and SNPs.

The model identified a large number of significant FCs for the low WRAT (Wide Range Assessment Test) group. In contrast, for the high WRAT group, the model identified a smaller number of significant FCs. The authors used a hypothetical validation technique, which has little empirical significance. This study used ConsensusPathDB-human (CPDB) database as a reference to validate the identified SNPs. The authors provided probable explanations for the identified SNPs, clarifying the model’s discriminative behavior.

Lin et al. [164] proposed a 3D-CNN model to classify schizophrenia patients from normal controls using spatial source phase (SSP) maps derived from complex-valued fMRI data. This study showed the superior performance of SSP maps compared to magnitude maps (MAG) extracted from magnitude-only fMRI data, and spatial source magnitude maps (SSM) separated from complex-valued fMRI data. The authors used two interpretability methods, saliency maps and Grad-CAM, to separately understand the prominent and predictive regions associated with the model predictions. While CNN can be a powerful tool for feature extraction and classification, the underlying caveat was the susceptibility of model performance and associate heatmaps because they varied widely according to the number of convolutional layers used.
Zhang et al. [22] proposed a learning framework combining the residual network and self-attention to perform two classification tasks using sMRI images: classifying AD from NC and pMCI from sMCI. This study, in particular, showed that residual networks could learn from sMRI images compared to other variants of convolutional networks (e.g., 3D-VGGNet) and self-attention helps to upgrade the classification performance. The authors applied 3D Grad-CAM to explain individual predictions. One problem with Grad-CAM in understanding the characteristic patterns responsible for predictions is that it cannot capture the fine details in the brain space because of required upsampling. Often, people use convolution layers close to the input layer to increase the resolutions of heatmaps. However, different convolution layers learn different levels of abstraction from the data. So, in that case, explanation maps may not reflect the global behavior of the model.

Leming et al. [218] used a diverse collection of fMRI datasets and leveraged a deep convolutional neural network for three different classification tasks—ASD, gender, and resting/tasks—using functional connectivity (FC). The authors showed that the deep learning model is capable of good classification when datasets are a mixture of multi-site collections. The authors used the 116-area automated anatomical labeling (AAL) parcellation template [219] and computed functional connectivity of $4 \times 116 \times 116$ (4 wavelet frequency scales and 116 nodes wavelet coefficient correlation). This study showed that CAM could identify the brain’s prominent spatial elements (connectome) that the models used for predictions. In contrast, activation maximization, though initially used to gain intuitions of neural network internals [220], was able to provide insights into the critical predictive features suitable for
classification. However, as the variation of the accuracies of the ensemble was very large, the identified areas may not fully characterize ASD.

**Gradients and Guided Backpropagation**

Rieke et al. [221] proposed a 3D-CNN to classify AD patients from healthy controls. The authors used four visualization methods—gradients, guided backpropagation, occlusion, and brain area occlusion—to generate explanations. Relevance scores from gradient-based visualization methods were more distributed across the brain, as opposed to occlusion and brain area occlusion, where relevance scores are more focused on specific regions. Distributive relevance is not feasible for occlusion-based methods because of the limited size of the patch. Hence, the authors recommend using gradient-based approaches for scenarios where distributed relevance is expected. Unlike LRP, as claimed in [222], the authors think that similar heatmaps for both AD and NC are reasonable because a given network should look into similar regions to detect the absence or presence of the disease. To quantify the difference between visualization methods, the authors used Euclidean distance between average heatmaps of the groups (AD or HC) obtained from two visualization methods. Gradient-based methods showed a very small distance.

Oh, et al. [27] proposed a CNN-based end-to-end learning model to perform four different classification tasks classifying various stages of AD (Alzheimer’s disease) from NC (normal control) and pMCI from sMCI. The study used a convolutional autoencoder to pretrain the model in an unsupervised fashion. The authors, after prediction, used the saliency method (gradients) to visualize predictive features that the models used for each classification. Analysis
of the heatmaps revealed that the temporal and parietal lobes were most discriminative between AD patients and controls.

**Layer-wise Relevance Propagation**

DeepLight \[134\] proposed a DL model consisting of recurrent (LSTM) and convolutional elements to analyze the whole-brain activity associated with cognitive states. Each whole-brain volume is sliced into a set of axial images to feed into the convolutional and recurrent units. To generate post hoc explanations, DeepLight uses LRP (Layer-wise Relevance Propagation) \[54\]. The model was trained to predict four different cognitive states corresponding to four stimulus classes (body parts, faces, places, or tools). The baselines used to assess the effectiveness were General Linear Model, Searchlight Analysis, and Whole-Brain Least Absolute Shrinkage Logistic Regression. The model takes each brain volume and then passes through a combination of convolutional and recurrent DL elements to predict the volume corresponding cognitive state. Along the time dimension, it produces a sequence of predictions, one for each sample time point. LSTM here is indeed used for learning spatial dependency within and across the brain slices. After each prediction for each brain volume, the LRP method is used to generate a post hoc explanation for that prediction attributing relevance to the voxel levels. LRP was used only for the correct predictions. The overall accuracy was around 68.3% on the held-out dataset. The validation or evaluation of the quality of the maps was achieved through a meta-analysis of the four cognitive states using an established cognitive state-brain association database called NeuroSynth. Hu et al. \[223\] proposed a deep learning framework, called *Deep Collaborative Learning (DCL)*, for efficient integration of
different data modalities. The authors show that deep models that integrate multimodal data can better learn complex non-linear relationships from the data.

Eitel et al. [216] investigated the possibility of layer-wise relevance propagation (LRP) to uncover the rationale behind decisions made by 3D convolutional neural networks (CNNs) trained to diagnose multiple sclerosis (MS). The identified features revealed that CNN, in conjunction with LRP, has the potential to identify relevant imaging biomarkers, for example, individual lesions, lesion location, non-lesional white matter, or gray matter areas. These biomarkers are considered established MRI markers in MS literature.

Bohle et al. [222] used LRP to explain the decisions of a CNN model. They used a scalable brain atlas [224] and defined two metrics, "relevance density" and "relevance gain," for objective assessment of the heatmaps. The key reason behind using LRP rather than gradient-based methods is that LRP decomposes the output in terms of contributions in the input space. As the authors mentioned, LRP has the potential to answer this question — "what speaks for AD in this particular patient?" where explanations using gradient-based approaches apparently address the following question: "which change in voxels would change the outcome most?" We argue that these two questions are not mutually exclusive. For a comparison of LRP with gradient-based methods, the authors used "guided-backpropagation." While both LRP and GB were successful in localizing important regions, GB, compared to LRP, showed less contrast in importance scores between group-wise (AD vs. HCs) heatmaps. Fortunately, there are other gradient-based methods (e.g., integrated gradients [43] and smoothgrad [44] on integrated gradients) with desirable properties that future studies may
Several studies have attempted to learn from different modalities. For example, Zhao et al. [198] proposed a hybrid deep learning architecture to combine sequential temporal dynamics (TCs) and functional dependency (FNCs). The authors used an attention module on top of C-RNN to extract temporal dynamic dependencies from TCs and used LRP to identify the most group-discriminative FNC patterns. Please note that LRP was used in a post hoc manner for the analysis of FNC patterns, not as part of the learning process.

Hofmann et al. [188] proposed ensembles of convolutional neural networks with LRP to identify which neural features contribute most to brain age. The models were acceptably accurate and could capture aging at both small and large-scale changes. The models were also able to identify associated risk factors in case of diverging brain age. The study detected three major brain components (gray matter, white matter, and cortical spinal fluids) whose relevance scores were linearly correlated to the function of age. The authors argued in favor of ensemble models because the variability of predictions between different models, even when they have the same architecture and are trained on the same data, may arise because of the high variance and bias of individual models. Multiple studies have recommended aggregation of saliency maps generated from single base models [188, 225]. LRP, similar to other prevailing explanation methods, cannot inform us anything about the underlying biological mechanisms justifiable for the generated explanations.

### Integrated Gradients and Smoothgrad

In a recent study, Rahman et al. [18] proposed an interpretable deep learning framework.
The framework includes a pre-trainable model suitable for multiple downstream studies with limited data size. The authors also proposed how we can investigate spatio-temporal dynamics associated with mental disorders using post hoc interpretability methods (integrated gradients (IG) and smoothgrad on integrated gradients). Apart from qualitative evaluation, the framework suggested a quantitative evaluation technique, called RAR, to objectively show that identified salient regions are indeed meaningful and highly predictive. This study demonstrates the utility of IG and smoothgrad for neuroimaging interpretability.

Levakov et al. [225] proposed an ensemble of CNNs and aggregate "explanation maps" to arrive at some conclusive remarks associated with brain age. The authors used smoothgrad as a post hoc interpretability method and were particularly interested in population-wise explanation rather than subject-specific identification of anatomical brain regions. This study also used ensembles of CNN to analyze the model uncertainty behavior. Population-based map for each ensemble was produced by averaging all the volumes in the test set. To generate the global population-based map, they aggregate population-based maps generated for each CNN by taking the median value for each voxel across the ensembles. While this approach highlights important areas in the brain space, this approach is not able to comment on the direction of influence. It is impossible to determine if the regions contribute positively or negatively to brain age.

Wang et al. [168] applied Integrated Gradients (IG), LRP, and Guided Grad CAM to visualize CNN models designed for Alzheimer's classification. The authors observed that IG is the best, as revealed in the meta-analysis performed on top of all visualizations. IG
heatmaps were particularly more focused on the hippocampus than Guided Grad-CAM and LRP heatmaps, consistent with well-supported biomarkers for Alzheimer’s disease.

Zeineldin et al. [165] compared seven popular gradient-based explanation methods: gradients, Smoothgrad, integrated gradients, guided backpropagation (GBP), gradient-weighted class activation map (Grad-CAM), Guided Grad-CAM, and Guided Integrated Gradients for MRI image classification and segmentation tasks. For the classification task, Guided Grad-CAM (i.e., combining GBP with GCAM) produced better localization, while Smoothgrad provided the best discriminative regions of the input. For the segmentation task, Smoothgrad was found to be the best choice because of its robustness to noise, while GCAM did the best visualization as it identified the most discriminative regions.

### 3.5.2 Perturbation-based Methods

**Occlusion Sensitivity**

Abrol et al. [157] experimented with a modified deep ResNet to predict the progression to AD. While the main focus was to predict the progression from MCI class to AD class, the study also experimented with eight combinations of binary, mixed-class (based on transfer learning), and multi-class diagnostic and prognostic tasks. The authors also leveraged network occlusion sensitivity to identify the anatomical regions that were most predictive for the progression of MCI to AD. In the analysis, thirteen brain regions, including the middle temporal gyrus, cerebellum crus 1, precuneus, lingual gyrus, and calcarine, consistently emerged in the top 20 most relevant regions. As the occlusion sensitivity method considers only the output score drop due to occlusion of a defined region and does not consider connectivity among regions,
the method does suffer from several limitations as pointed out by [27, 176]. The authors also projected the features from the first fully-connected layer onto a 2-dimensional space using t-SNE (t-distributed stochastic neighbor embedding) to demonstrate the separability of the learned representations.

3.5.3 Distillation Methods

Local Interpretable Model-agnostic Explanations (LIME)

Magesh et al. [190] leveraged the VGG16 network pretrained on ImageNet dataset to classify Parkinson’s patients from healthy controls. The study also used LIME to explain individual predictions. The underlying reason for choosing this explanation method is unclear. Furthermore, while quantitative validation is an essential measure of the predictability of the heatmaps, the study did not conduct any experiments to validate the generated explanations objectively.
Table 3.2: Literature Review of Interpretable Deep Learning Research in Neuroimaging

<table>
<thead>
<tr>
<th>Authors, Year</th>
<th>Study Objective</th>
<th>Dataset</th>
<th>Modality</th>
<th>DL Component</th>
<th>Accuracy</th>
<th>Interpretability</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang et al., 2018 [176]</td>
<td>AD Classification</td>
<td>ADNI</td>
<td>sMRI</td>
<td>VGGNet, ResNet, ResNet-GAP, Shallow-GAP</td>
<td>0.86 ± 0.05, 0.85 ± 0.08, 0.64 ± 0.11, 0.75 ± 0.08</td>
<td>Occlusion, SA-3DUCM, 3D-CAM, 3D-GRAD-CAM</td>
<td>previous hypotheses</td>
</tr>
<tr>
<td>Rieke et al., 2018 [221]</td>
<td>AD Classification</td>
<td>ADNI</td>
<td>sMRI</td>
<td>3D-CNN</td>
<td>0.77 ± 0.06</td>
<td>Gradients, GB, Occlusion, Brain Area Occlusion</td>
<td>AAL atlas quantitative</td>
</tr>
<tr>
<td>Thomas et al., 2019 [134]</td>
<td>cognitive state prediction</td>
<td>HCP</td>
<td>rsfMRI</td>
<td>bi-LSTM CNN</td>
<td>68.3%</td>
<td>( \epsilon )-LRP</td>
<td>meta analysis (NeuroSynth database)</td>
</tr>
<tr>
<td>Eitel et al., 2019 [216]</td>
<td>Multiple Sclerosis Classification</td>
<td>ADNI</td>
<td>sMRI</td>
<td>3D-CNN</td>
<td>87.04%</td>
<td>( \epsilon )-LRP</td>
<td>previous hypotheses</td>
</tr>
<tr>
<td>Bohle et al., 2019 [222]</td>
<td>AD Classification</td>
<td>ADNI</td>
<td>sMRI</td>
<td>3D-CNN</td>
<td>87.96%</td>
<td>LRP-( \beta )</td>
<td>scalable atlas [224] based metrics</td>
</tr>
<tr>
<td>Authors, Year</td>
<td>Study Objective</td>
<td>Dataset</td>
<td>Modality</td>
<td>DL Component</td>
<td>Accuracy</td>
<td>Interpretability</td>
<td>Validation</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
<td>---------</td>
<td>----------</td>
<td>--------------</td>
<td>----------</td>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>Rahman et al., 2022 [18]</td>
<td>SZ AD ASD Classification</td>
<td>FBIRN OASIS ABIDE</td>
<td>rsfMRI</td>
<td>LSTM Attention</td>
<td>77% 70% 63%</td>
<td>IG Smoothgrad-IG</td>
<td>RAR framework, previous hypotheses</td>
</tr>
<tr>
<td>Zhao et al., 2022 [198]</td>
<td>SZ ASD Classification</td>
<td>In-House ABIDE</td>
<td>rsfMRI</td>
<td>C-RNN\textsuperscript{AM} DNN</td>
<td>85.3% 72.4%</td>
<td>LRP</td>
<td>previous hypotheses</td>
</tr>
<tr>
<td>Hu et al., 2021 [217]</td>
<td>LOW/HIGH WRAT Classification</td>
<td>PNC Illumina HumanHap Illumina Human Omni</td>
<td>nback-fMRI genomic data</td>
<td>ConvNets</td>
<td>75.01%</td>
<td>Grad-CAM GBP</td>
<td>previous hypotheses, Gene enrichment analysis (CPDB) database</td>
</tr>
<tr>
<td>Chen et al., 2022 [175]</td>
<td>ASD Classification</td>
<td>ABIDE sMRI</td>
<td>3D-ResNet, attention subnet</td>
<td>75%</td>
<td>Grad-CAM</td>
<td>previous hypotheses</td>
<td></td>
</tr>
<tr>
<td>Yan et al., 2017 [226]</td>
<td>SZ Classification</td>
<td>Chinese rsFMRI</td>
<td>DNN</td>
<td>82%</td>
<td>LRP</td>
<td>previous hypotheses</td>
<td></td>
</tr>
<tr>
<td>Authors, Year</td>
<td>Study Objective</td>
<td>Dataset</td>
<td>Modality</td>
<td>DL Component</td>
<td>Accuracy</td>
<td>Interpretability</td>
<td>Validation</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------</td>
<td>---------</td>
<td>----------</td>
<td>--------------</td>
<td>----------</td>
<td>------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Levakov et al., 2020 [225]</td>
<td>Age Prediction</td>
<td>collected from 15 open databases</td>
<td>T1w-MRI</td>
<td>Ensembles CNN</td>
<td>MAE 3.07 years</td>
<td>Smoothgrad</td>
<td>Replicability, Similarity, Specificity, tests</td>
</tr>
<tr>
<td>Hofmann et al., 2022 [188]</td>
<td>Brain Age Estimation</td>
<td>LIFE Adult Study</td>
<td>sMRI (T1, FLAIR, SWI)</td>
<td>Ensembles CNNs</td>
<td>MAE 3.37–3.86 years</td>
<td>LRP</td>
<td>simulation study, brain atlases, significance tests</td>
</tr>
<tr>
<td>Lin et al., 2022 [164]</td>
<td>SZ Classification</td>
<td>UNM IRB complex-valued rsfMRI</td>
<td>3D-CNN</td>
<td>sMOT/90.8%, DMN/96.0%, AUD/98.4%</td>
<td>Gradients</td>
<td>Grad-CAM</td>
<td>previous hypotheses</td>
</tr>
<tr>
<td>Magesh et al., 2020 [190]</td>
<td>PD Classification</td>
<td>PPMI SPECT DaTSCAN</td>
<td>VGG16 (CNN)</td>
<td>95.2%</td>
<td>LIME</td>
<td>previous hypothesis</td>
<td></td>
</tr>
<tr>
<td>Zhang et al., 2021 [22]</td>
<td>AD v. NC pMCI v. sMCI Classification</td>
<td>ADNI-1 ADNI-2 ADNI-3</td>
<td>sMRI ResAttNet34 (3D)</td>
<td>0.913 0.821</td>
<td>Grad-CAM</td>
<td>previous hypotheses</td>
<td></td>
</tr>
<tr>
<td>Authors, Year</td>
<td>Study Objective</td>
<td>Dataset</td>
<td>Modality</td>
<td>DL Component</td>
<td>Accuracy</td>
<td>Interpretability</td>
<td>Validation</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>---------</td>
<td>----------</td>
<td>--------------</td>
<td>----------</td>
<td>------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Oh et al., 2019 [27]</td>
<td>AD vs. NC</td>
<td>ADNI</td>
<td>sMRI</td>
<td>CAE Inception</td>
<td>AD-86.60%</td>
<td>Gradients</td>
<td>previous hypotheses</td>
</tr>
<tr>
<td></td>
<td>pMCI vs. NC</td>
<td></td>
<td></td>
<td></td>
<td>pMCI-73.95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sMCI vs. NC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pMCI vs. sMCI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abrol et al., 2020 [157]</td>
<td>CN vs. AD</td>
<td>ADNI</td>
<td>sMRI</td>
<td>ResNet SAE</td>
<td>89.3%</td>
<td>FM visualization</td>
<td>previous reports</td>
</tr>
<tr>
<td></td>
<td>CN vs. pMCI</td>
<td></td>
<td></td>
<td></td>
<td>86.5%</td>
<td>Occlusion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sMCI vs. AD</td>
<td></td>
<td></td>
<td></td>
<td>87.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sMCI vs. pMCI</td>
<td></td>
<td></td>
<td></td>
<td>75.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biffi et al., 2020 [227]</td>
<td>HCM vs. NC</td>
<td>multi-site cohort &amp; ADNI</td>
<td>sMRI</td>
<td>generative models</td>
<td>100%84%</td>
<td>FM visualization</td>
<td>previous hypotheses</td>
</tr>
<tr>
<td></td>
<td>AD vs. NC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martinez-Murcia et al., 2019 [228]</td>
<td>AD</td>
<td>ADNI</td>
<td>sMRI</td>
<td>CNN autoencoders</td>
<td>84%</td>
<td>FM visualization</td>
<td>previous hypotheses</td>
</tr>
<tr>
<td></td>
<td>Classification, predicting other variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authors, Year</td>
<td>Study Objective</td>
<td>Dataset</td>
<td>Modality</td>
<td>DL Component</td>
<td>Accuracy</td>
<td>Interpretability</td>
<td>Validation</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>---------</td>
<td>----------</td>
<td>--------------</td>
<td>----------</td>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>Leming et al., 2020 [218]</td>
<td>ASD vs. TD Gender</td>
<td>1000FC</td>
<td>ABCD</td>
<td>ASD vs. TD Gender Task vs. Rest Classification</td>
<td>0.6774</td>
<td>AM prior findings</td>
<td>Grad-CAM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABIDE</td>
<td></td>
<td></td>
<td>0.7680</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ABIDE II</td>
<td>fMRI</td>
<td>CNN</td>
<td>0.9222</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADNI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>UKBiobank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NDAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>OpenfMRI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.5.4 Feature Map Visualization

Biffi et al. [227] proposed a hierarchical deep generative model called ladder variational autoencoder (LVAE). LVAE learns a hierarchy of conditional latent variables to represent the population of anatomical segmentations. The latent space representation in the highest level of the hierarchy can efficiently discriminate clinical conditions. The proposed model performed two classification tasks: 1) Hypertrophic cardiomyopathy (HCM) versus healthy 3D left ventricular (LV) segmentations and 2) AD versus healthy control 3D hippocampal segmentations. The model was predictive of clinical conditions and offered suitable visualization and quantification of the anatomical shape changes associated with those clinical conditions. This study used sampling in the highest latent space to visualize the corresponding regions in the brain space. The authors further claimed that the shape changes, as evident in the visualization, agreed with the clinical literature.

Martinez-Murcia et al. [228] used a deep CNN autoencoder for an exploratory data analysis of AD. The autoencoder demonstrates links between cognitive symptoms and the underlying neurodegenerative process. The autoencoder model uses a data-driven approach to extract imaging characteristics into low-dimensional manifolds. The study further used regression analysis to show that the neurons in the manifold space correlate well with the clinical and neuropsychological test outcomes and diagnoses. Subsequently, the authors used a novel visualization approach using a linear decomposition model to show the brain regions highly influenced by each manifold coordinate, which provides additional information about the association between structural degeneration and the cognitive decline of dementia.
3.5.5 Intrinsic Methods

Very few neuroimaging studies so far have considered interpretability as part of the algorithmic aspect of the model from its inception. Such models in the literature are called glass-box or transparent-box models. Biffi et al. [227] proposed a deep generative model for transparent visualization of the classification space. Some other neuroimaging studies [210, 229] considered interpretable models based on their design transparency.

3.6 The Usage Trend of Interpretability Methods

While the neuroimaging community has used a larger collection of interpretability methods, only a few are popular and considered important for knowledge discovery or potential clinical deployment. Several interpretability methods have often been used as experimental baselines, not for their beneficial effects in this domain. In this section, we conducted an in-depth analysis of the usage of all popular interpretability methods in neuroimaging studies. Indeed, we investigated the usage of these methods in more than 300 neuroimaging papers and observed their usage trend as shown in Table 3.3. As we found in our exploratory analysis, studies have used the methods in the following order of frequency: 1) CAM/Grad-CAM/Grad-CAM++/Guided Grad-CAM [50, 87, 88] 2) SHapley Additive exPlanations (SHAP) [47] 3) Integrated Gradients [43] 4) Layer-wise Relevance Propagation [54] 5) Occlusion Sensitivity [48] 6) Guided Backpropagation [51] 7) Local Interpretable Model-Agnostic Explanations (LIME) [49] 8) Gradients [52, 53] 9) DeepLIFT [55] and 10) Smoothgrad [44].

This usage trend also reveals that preference for "gradients" and "guided backpropagation"
methods are receiving less attention because of their limitations [44, 109], while the steeper rise in the use of integrated gradients and SHAP are potentially due to their strong theoretical foundations.

Table 3.3: The usage trend of popular post hoc interpretability methods.

<table>
<thead>
<tr>
<th>Interpretability Method &amp; Citing Publications</th>
<th>Usage Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradients [52, 53]: [27, 164–167, 174, 181, 205, 221, 264, 272–279]</td>
<td>![Graph]</td>
</tr>
</tbody>
</table>
Table 3.3: The usage trend of popular post hoc interpretability methods.

<table>
<thead>
<tr>
<th>Interpretability Method &amp; Citing Publications</th>
<th>Usage Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Integrated Gradients</strong> [43]: [18, 158, 159, 165–172, 178, 184, 211, 241, 305, 411–429]</td>
<td></td>
</tr>
<tr>
<td><strong>Smoothgrad</strong> [44]: [18, 165, 166, 174, 177, 273, 430, 431]</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3: The usage trend of popular post hoc interpretability methods.

<table>
<thead>
<tr>
<th>Interpretably Method &amp; Citing Publications</th>
<th>Usage Trend</th>
</tr>
</thead>
</table>
Table 3.3: The usage trend of popular post hoc interpretability methods.

<table>
<thead>
<tr>
<th>Interpretability Method &amp; Citing Publications</th>
<th>Usage Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Interpretable Model-Agnostic Explanations (LIME) [49]: [190–194, 292, 313, 468, 478, 480, 507, 512, 522–530]</td>
<td><img src="image" alt="Graph showing the usage trend of popular post hoc interpretability methods." /></td>
</tr>
</tbody>
</table>

3.7 Suggestions for Interpretable Models in Neuroimaging

In this section, we discuss the significant pitfalls of interpretability research in neuroimaging. One of the obvious concerns in interpretable deep learning models is stability in explanations. As such, explanations vary widely among architectures and interpretability methods. As different neural networks may assign different regions as important for predictions, most of them would tell about different aspects of the disorder. Combining explanations from different models and further analysis of these explanations in association with medical experts may be useful in revealing undiscovered aspects of the disease. To this end, a unified framework [156] in interpretable neuroimaging research may be useful so that the findings across the studies can be directly compared to communicate advancement benchmarks. Based on our analysis and review, we recommend that everyone focus on the following directions: 1) Objective
quantification of the performance, 2) Investigation of the sensitivity to interpretability parameters, 3) Investigation of causality or underlying mechanism on top of heatmaps, 4) Investigating the reliability of the underlying model via model debugging 5) Even when models use "true" evidence, explanations and their relative importance may be different, further analysis for combining these different aspects is required. As Rieke et al. [221] pointed out that different visualization methods (gradients or non-gradient approaches) vary widely, so in line with other earlier studies, we suggest investigating multiple methods instead of blindly relying on one method, given the interpretation task of a model.

While many earlier studies used the occlusion sensitivity method to generate explanations, Yang et al., 2018 [176] pointed out several limitations of the approach. For example, this approach uses semantically meaningless neighborhoods and an unspecified way of choosing the grid size. Moreover, the method is computationally very intensive. As no backpropagation from the target score is involved during heatmap generation, this explanation is considered to be limited [27].

Also, 3D-Grad-CAM can be useful if we need to track the attention of the convolution layers, but Grad-CAM or CAM is not useful for generating explanations in the input space and hence not suitable for data interpretation. Though LRP has been used extensively, it has inherent limitations. LRP cannot maintain implementation invariance as it uses modified backpropagation rules. For future interpretability practices, we leave the following suggestions for the neuroimaging community:

1. **Post hoc methods are blamed for being insufficient:** As post hoc methods
heavily rely on the models they are applied to, the methods can only discover the minimal discriminative parts sufficient for the prediction. For example, while LRP and GBP have been shown to be able to identify homogeneous brain regions, e.g., the hippocampus, they cannot identify heterogeneous regions, e.g., cortical folds [161, 222].

2. **Lack of any guiding principle to select explanation methods:** While studies have leveraged different explanation methods for deep learning models, there is little theoretical evidence or guiding principle to choose a method for a particular study. Recently, Han et al. [46] demonstrated how different explanation methods describe different neighborhoods and thus produce different explanations. Some disagreement scenarios are common because there could be differences in the underlying aspects the methods are investigating. For example, permutation importance [531] and SHAP [47, 532] in case of model overfitting may produce very different explanations. However, some disagreement scenarios are not expected. For example, gradients and LIME should produce similar interpretations because they both focus on local neighborhoods. However, in practice, they produce very different explanations. The authors in [46] also showed how some methods cannot recover the underlying model and are entirely independent. The authors also provided valuable suggestions on choosing interpretability methods based on the nature of the data. They further suggested building an explanation method for the data for which no explanation method from the literature is considered beneficial.

3. **We need to be aware of the fragility of neural network interpretations:** The
fundamental problem with the popular interpretability methods is their robustness [533]. Ghorbani et al. [533] showed that interpretations based on feature importance maps such as DeepLIFT, integrated gradients, and influence functions are susceptible to adversarial attacks. Put another way, a systematic perturbation of the input can lead to a very different interpretation (heatmap) because of the complexity of input feature space in deep neural networks. In neuroimaging, earlier studies, so far we are aware, usually overlooked this fragility of the interpretations, which may lead to misleading interpretations. Moreover, there is an inherent human bias to trust the model as correct and look for interpretations only based on predictive performance. While model inspection or debugging can be a hard problem in neuroimaging, it should be an essential consideration for this safety-critical domain.

4. Attributions normalization and polarity considerations varied widely: For the post-processing of different explanations, studies use an ad-hoc approach. There has yet to be an agreement on how to post-process the heatmaps. This agreement must correspond to the underlying model and the interpretability method used. This necessity of the agreement is especially applicable to gradient-based attribution methods. Studies used the sign information differently to finalize the heatmaps.

5. Studies generally use an ad-hoc approach to validate explanations: For the validation of results, studies generally use informal and unreliable ways. Sometimes they used intuitions, hypotheses, and earlier results to justify the current attributions. These validation techniques are very susceptible and may end up with misleading
conclusions. As Levakov et al. [225] indicated, any reasonable conclusions regarding the contributions should be made based on common parts of the maps from multiple models. Furthermore, deep learning models usually capture complex hierarchical and multivariate interactions. Localizing the brain regions should only be considered as an approximation of the significance. Even a small architectural modification can be a significant determinant of model performance and feature attribution maps, as indicated by Lin et al. [164].

6. Predictability of explanations may not be sufficient: While RAR/ROAR-based evaluation of the salient regions is promising and may further enhance the trust in the significance of what the model has learned, it may still need to be guaranteed that the model did not rely on spurious correlations. The domain experts should confirm the validation of the interpretations. Equivalently the explanations must match a significant proportion of the expert-extracted knowledge. We suggest complementing quantitative validation with neurologically valid explanations.

7. Use structure-function fusion model for model diagnosis: Earlier studies, in general, independently focused on the anatomical or functional aspects of the dynamics. However, using both modalities simultaneously and corresponding existing knowledge in each modality during explanation generation may provide rigorous validation and bring trust in the explanations.

8. Counterfactual explanations may reveal the underlying biological mecha-
nism: Wachter et al. [79] first introduced counterfactual explanations to know about the hypothetical reality that could alter the model’s decision. Dandi et al. [81] refined the formulation to satisfy the different practical desiderata of counterfactual explanations to make them useful in real-world applications. In the context of neuroimaging, we believe counterfactual explanations may help understand the underlying biological mechanism that potentially caused the specific disorder in the first place. To our knowledge, no neuroimaging study has ever used counterfactuals to understand the model’s decision-making process.

9. Layer-wise Relevance Propagation (LRP) needs further investigation: As seen from the interpretability in neuroimaging literature, LRP has been widely used, and its popularity is on an upward trend. However, the explanations produced by LRP are not reliable. Indeed, Shrikumar et al. [55] showed a strong connection between LRP and \( \text{gradient} \odot \text{input} \), especially when all the activations are piecewise linear as in ReLU or Leaky ReLU. Ancona et al. [78] also showed that \( \epsilon \)-LRP is equivalent to the feature-wise product of the input and the modified partial derivative. Kindermans et al. [534] showed that DeConvNet, Guided BackProp, and LRP cannot produce the theoretically correct explanation even for a linear model—the most straightforward neural network.

10. SHAP is popular, but it should not be trusted blindly: SHAP, though very popular in the XAI community, has some issues. For example, SHAP assumes that the features are independent, while they are very unlikely. While features may be correlated,
the algorithm may generate unrealistic observations (instances) with permutations. Moreover, no explanation method produces explanations that imply causality. SHAP indicates the importance of a feature based on the model prediction, not the importance in the real world. Humans are very prone to confirmation bias. It is not very uncommon that humans tend to create narratives as a result of confirmation bias. The most important question is: Did the model learn to predict for the right reasons? This question is vital because machine learning models do not know about truths, and it only cares about correlations, and proxy or secondary or less important variables may be loosely or tightly correlated with the actual cause. They can be revealed as very important features. Moreover, Kwon and Zou [535] recently showed that SHAP is suboptimal in that it gives the same weight to all marginal contributions for a feature $x_i$, which may potentially lead to attribution mistakes if different marginal contributions have different signal and noise. The authors further proposed a simple modification of the original SHAP, called WeightedSHAP, that estimates the weights automatically from the data.

11. **Studies generally focused only on classification and regression tasks:** While many studies in interpretable deep learning models for general classification tasks exist, further subgrouping into patient subtypes or clustering is still a novel area. This lack of interpretability literature for clustering tasks is equally true for neuroimaging and other domains. Very few studies did projection transformation from the latent space to observe the area of influence [227, 228].
12. Effectiveness of transfer learning in neuroimaging needs justification: what causes the increased accuracy? What knowledge does it transfer? Raghu et al. [36] showed that transfer learning from natural images to medical images did help little with performance. Instead, as the authors surmised, the slight improvement may come from the over-parameterization of the standard models trained on natural images. Moreover, studies are not certain about the aspects of knowledge they are transferring from the natural image domain to the medical image domain or from one disorder area to another.

3.8 Conclusion

In this chapter, we discuss the significance of interpretability in neuroimaging studies that were built upon deep learning approaches. We reviewed more than 300 neuroimaging studies that considered model interpretability as their essential component. Moreover, we presented example neuroimaging studies for all the prevailing interpretability methods. We reckon that these analyses will be helpful for future neuroimaging practitioners looking for a general guideline. Additionally, we analyzed the recent usage trend of the most prevailing post hoc interpretability methods, which clearly shows their continued acceptance in the neuroimaging community. Finally, we discuss different caveats of interpretability practices and provide insights on how this specialized sub-field of AI can be used wisely and meaningfully.
CHAPTER 4
TRANSFER LEARNING FOR NEUROIMAGING

The traditional voxel-based analysis has significant limitations, and scientists have been using AI-based approaches for the last few decades [6, 8]. However, standard machine learning models (SML) have limitations when they handle high-dimensional data, for which applying feature reduction techniques becomes an obvious step. These limitations necessitate the use of DL models to learn directly from data [23]. However, training reliable DL models is a challenging step in neuroimaging due to the traditional data scarcity problem in the domain. In this chapter, we discuss how we can formulate a self-supervised contrastive pre-training method so that the model can gain foundational knowledge from an unrelated and unlabeled dataset, which eventually improves the downstream model performance both in synthetic and real datasets.

4.1 Introduction

Mental disorders manifest in behavior that is driven by disruptions in brain dynamics. Functional MRI captures the nuances of spatiotemporal dynamics that could potentially provide clues to the causes of mental disorders and enable early diagnosis. However, the obtained data for a single subject is of high dimensionality $m$, and to be useful for learning, and statistical analysis, one needs to collect datasets with a large number of subjects $n$. Yet, for any disorder, demographics, or other types of conditions, a single study is rarely able to amass datasets large enough to go out of the $m >> n$ mode. Traditionally this is approached by handcrafting features [15] of a much smaller dimension, effectively reducing $m$
via dimensionality reduction. Often, the dynamics of brain function in these representations vanishes into proxy features such as correlation matrices of functional network connectivity (FNC) [226]. Efforts that pull together data from various studies and increase $n$ do exist, but it is difficult to generalize to the study of smaller and more specific disease populations that cannot be shared to become a part of these pools or are too different from the data in them.

Our goal is to enable the direct study of brain dynamics in smaller datasets to, in turn, allow an analysis of brain function. In this paper, we show how one can achieve significant improvement in classification directly from dynamical data on small datasets by taking advantage of publicly available large but unrelated datasets. We demonstrate that it is possible to train a model in a self-supervised manner on the dynamics of healthy control subjects from the Human Connectome Project (HCP) [536] and apply that pre-trained encoder to a completely different data collected across multiple sites from healthy controls and schizophrenia subjects.

4.2 Related Work

Recent advances in unsupervised learning using self-supervised methods by estimating and maximizing mutual information reduced the gap between supervised and unsupervised learning [537–539]. Such success has already influenced neuroimaging in the case of structural MRI [540] and even reinforcement learning [541].

Prior works in brain imaging have been based on unsupervised methods such as linear ICA [542] and HMM framework [543]. Some other nonlinear approaches were also proposed
to capture the dynamics as using RBMs [544] and RNN modification of ICA [545].

Also, in most cases, researchers in brain imaging are dealing with small datasets. In this case, transfer learning [30, 31, 546] might be a way to improve results and in some cases, to enable learning from data otherwise too small for any results. Another idea to improve performance might be considered by a data generating approach [547].

4.3 Method Description

For self-supervised pre-training, we are using spatio-temporal objective ST-DIM [541] to maximize predictability between current latent state and future spatial state and between consecutive spatial states. For the lower bound of mutual information, we are using InfoNCE [537] estimator. Compare to other available estimators, InfoNCE shows better performance [538, 539] in case of a greater number of negative samples that are readily available in case of time series data.

Let \( \{(u_t, v_s) : 1 \leq t, s \leq N, t \neq s\} \) be a dataset of pairs of values at time point \( t \) and \( s \) sampled from sequence with length \( N \). A pair \((u_t, v_{s+1})^+\) is called positive if \( s = t + 1 \) and \((y_t, v_s)^-\) — negative if \( s \neq t + 1 \). A positive pair models the joint and a negative — marginal distributions. Eventually, the InfoNCE estimator is defined as:

\[
J_f \left( \{(u_t, v_{t+1})^+\}_{t=1}^N \right) = \sum_{t=1}^{N} \log \frac{\exp f((u_t, v_{t+1})^+)}{\sum_{s=1}^{N} \exp f((u_t, v_s)^-)}, \quad (4.1)
\]

where \( f \) is a critic function [548]. Specifically, we are using separable critic \( f(u_t, v_s) = \phi(u_t)^T\psi(v_s) \), where \( \phi \) and \( \psi \) are some embedding function parametrized by neural networks. Such embedding functions are used to calculate value of a critic function in same dimensional
space from two dimensionally different inputs. Critic learns an embedding function such that it assigns higher values for positive pairs compare to negative pairs: $f((u_t, v_{t+1})^+) \gg f((u_t, v_s)^-)$.

We define a latent state as an output $z_t$ of encoder $E$ and a spatial state $c^l_t$ as the output of $l$th layer of the encoder for input $x_t$ at time point $t$. To optimize the objective between current latent state and future spatial state the critic function for input pair $(x_t, x_s)$ is $f_{LS} = \phi(z_t)^T\psi(c^l_t)$ and for consecutive spatial states $- f_{SS} = \psi(c^l_t)^T\psi(c^l_s)$. Finally, the loss is the sum of the InfoNCE with $f_{LS}$ and InfoNCE with $f_{SS}$ as $L = I_{f_{LS}} + I_{f_{SS}}$.

### 4.4 Experiments

#### 4.4.1 Simulation Data

To simulate the data we generate multiple 10-node graphs with $10 \times 10$ stable transition matrices. Using these we generated multivariate time series with autoregressive (VAR) and structural vector autoregressive (SVAR) models [549].

First, we generate 50 VAR times series with size $10 \times 20000$. Then we split our dataset to $50 \times 10 \times 14000$ samples for training, $50 \times 10 \times 4000$ — for validation and $50 \times 10 \times 2000$ — for testing. Using these samples we pre-train an encoder and evaluate based on its ability to identify consecutive $10 \times 20$ windows sampled from whole time series.

In the final downstream task we classify the whole time-series whether it is generated by VAR or SVAR (undersampled VAR at rate 2). We create 400 graphs with corresponding stable transition matrices and generate $2000 \times 10 \times 4000$ samples (5 for each) and split as
1600 × 10 × 4000 for training, 200 × 10 × 4000 for validation and 200 × 10 × 4000 for hold-out test. Here we also use 10 × 20 windows as a single time-point input.

4.4.2 Real data

Two independent datasets were used in the current study. The first dataset is a Schizophrenia dataset, which is from Function Biomedical Informatics Research Network (FBIRN)\(^1\), and the second dataset is a healthy subject dataset, which is from the 1200 Subject release of Human Connectome Project (HCP) \(^5\).

The FBIRN dataset was pre-processed through SPM12 \(^5\) based on the MATLAB 2016b environment. The slice-timing was first performed on the data, and then subject head motions were corrected by the realignment procedure. After that, the data was warped to MNI space using EPI template and resampled to 3 mm\(^3\) voxels. Finally, the data were smoothed with a 6mm FWHM Gaussian kernel. The FBIRN dataset consists of 311 subjects, including 151 SZ patients and 160 healthy controls.

The resting-state fMRI HCP data comes pre-processed by the following pipeline \(^5\). It includes removing of spatial distortions, compensation of the subject motion, reduction of the bias field, normalization with a global mean, and final brain masking. The pre-processed HCP data were then warped to the MNI space using the EPI template and resampled to the 3 mm\(^3\) voxels using the same bounding box, to guarantee HCP and FBIRN datasets have the same spatial resolution and dimensions. HCP consists only of healthy controls.

For each dataset, 53 intrinsic connectivity networks (ICNs) were extracted using the

---

\(^1\)These data were downloaded from the Function BIRN Data Repository, Project Accession Number 2007-BDR-6UHZ1
pipeline described at [553]. These 53 ICNs are supposed to be non-noise components providing meaningful functional network information and thus were used in training.

### 4.4.3 Training

Encoder for simulation experiment consist of 4 1D convolutional layers with output features \((32, 64, 128, 64)\), kernel sizes \((4, 4, 3, 2)\) and stride — 1, following by ReLU [554] after each layer followed by Linear layer with 256 units. For real data — 3 1D convolutional layers with \((64, 128, 200)\), \((4, 4, 3)\) and 1 respectively, followed by linear layer with 256 units. Then for all possible pairs in the batch we took flattened features after 3rd convolutional layer \(c^3\) and features from last layer \(z\). We embedded them using \(\psi\) for \(c^3\) and \(\phi\) for \(z\) to 128 dimensional vector to compute the score of a critic function \(f_{LS}\) or \(f_{SS}\). Using these scores we computed the loss. The neural networks trained using Adam optimizer [555]. The weights were initialized using Xavier [556].

For simulation experiment, first, we train our encoder to learn on 10 × 20 windows from the VAR time series using InfoNCE based loss, and secondly, we train a supervised classifier based on windows. This window-based classification provides promising results (accuracy 60%). However, in solving similar real problems, we are more interested in subjects, i.e., entire time series, rather than a single-window for classification. Hence, we perform classification based on the whole time-series. In this setting, the entire time-series is encoded as a sequence of representations and fed through a biLSTM classifier. Two additional linear layers with 200 hidden units on top of the last hidden state of the biLSTM are used to map the representation to classification scores.
For the real data case, similar to simulations, we successfully train (accuracy 87–90%) our encoder on consecutive windows of fMRI from HCP healthy subjects. Then each computed feature for each window of the whole fMRI sequence used to train biLSTM classifier on fBIRN dataset. The biLSTM classifies SZ and HC subjects. Overall each fMRI of the subject consists of a series of 13 overlapping by half windows by 53 components by 20 time points.

Figure 4.1: Left: Classification performance of schizophrenia patients from normal controls on the progressively larger datasets. Right: Simulated VAR vs. SVAR time-series classification performance. As we can observe, the ST-DIM pretraining improved the downstream discriminative performance for both the real and synthetic experiments.

4.4.4 Results

Here we compare an end-to-end supervised model without pre-training (NPT), with frozen layers of the pre-trained encoder (FPT), and with unfrozen layers of the pre-trained encoder (UFPT).

In the simulation study, we observe that the pre-trained model can easily be fine-tuned only with a small amount of downstream data. Our model can classify a randomly chosen
time-series as a sample of VAR or SVAR (refer to the right panel of Figure 4.1). Note, with very few training samples, models based on the pre-trained encoder outperform supervised models. However, as the number of samples grows, the accuracy achieved with or without pre-training levels out.

As we can see from the left panel of Figure 4.1, the real data results substantiate the insights achieved in a simulation study. The test dataset consists of 64 subjects that are held out from training and validation processes and are the same for all tests in the plot. Training data was randomly resampled ten times from the available data pool. To put it another way, self-supervised transferable pre-training always helps when we have very few samples offering higher AUC.

4.5 Conclusions and Future Work

As we have demonstrated, self-supervised pre-training of a spatiotemporal encoder on fMRI of healthy subjects provides benefits that transfer across datasets, collection sites, and to schizophrenia disease classification. Learning dynamics of fMRI helps to improve classification results for schizophrenia on small datasets, that otherwise do not provide reliable generalizations. Although the utility of this result is highly promising by itself, we conjecture that direct application to spatiotemporal data will warrant benefits beyond improved classification accuracy in the future work. Working with ICA components is smaller and thus easier to handle space that exhibits all dynamics of the signal. In the future, we will move beyond ICA pre-processing and work with fMRI data directly. We expect model introspection to yield
insight into the spatio-temporal biomarkers of schizophrenia. In future work, we will test the same analogously pre-trained encoder on datasets with various other mental disorders such as MCI and bipolar. We are optimistic about the outcome because the proposed pre-training is oblivious to the downstream use and is done in a manner quite different from the classifier’s work. It may indeed be learning crucial information about dynamics that might contain important clues into the nature of mental disorders.
CHAPTER 5
INTERPRETING BRAIN DYNAMICS

As DL models can learn better representations directly from the data, they have the potential to advance our understanding of domain-specific knowledge [37, 38]. However, interpreting the models and reliably extracting useful knowledge is very challenging because of the lack of reliable post hoc methods and a suitable validation technique. In this chapter, we discuss a complete interpretable DL framework, which supports improved pre-training (whole MILC) for fMRI data, provides reliable interpretations, and quantitative validation of the generated explanations for three mental disorders.

5.1 Introduction

Brain dynamics likely holds the key to understanding function and disorder [1–3]. The brain function manifests in a spatiotemporally localized activity within the dynamics [557]. Thus, identification and interpretation of subject-specific spatial and temporal activity may help guide our understanding of the disorder. Although, the spatiotemporal snapshots of brain dynamics can be captured noninvasively using functional magnetic resonance imaging (fMRI) [4, 5], the excessive dimensionality and complexity of fMRI signals rule out manual identification and interpretation. Alternatively, machine learning models trained to classify a mental disorder from the available observations have learned which aspects of the data reliably lead to correct prediction. In other words, the model builds internal representations of the mapping between the data and the class. Interpreting these representations can lead to discovery of previously unknown spatiotemporal functional indicators (or biomarkers).
However, standard machine learning (SML) models, when dealing directly with high-dimensional multivariate signals, suffer a drastic drop in performance because of the curse of dimensionality [17] (high dimensionality of fMRI relative to the typically available few samples). This is because the models are usually shallow and only learn simple relationships between input and output. To improve discriminative performance, neuroimaging researchers heavily rely on measures, such as cortical thickness or connectivity matrices [8, 558], that summarize spatiotemporal relationship between different brain regions [559, 560]. They apply some feature selection procedure on top of these measures to extract potentially useful features [15] to feed into the SML model. Arguably, such proxy, bias-prone representations rely on strict assumptions and miss the chance to discover highly predictive holistic representations of the underlying dynamics [561, 562]. Moreover, non-linear SML models are not easily interpretable.

Deep learning (DL) methods, on the other hand, are capable of learning complex hierarchical representations directly from the raw data through an increasingly higher level of abstraction. Recently, a large number of studies [8, 16, 24, 563] reported deep learning (DL) models’ potential in neuroimaging domains. For example, Abrol et al. (2021)[25] demonstrated the advantages of DL models trained on raw data over SML models trained on pre-engineered features in structural magnetic resonance imaging (sMRI). The study also suggests that the deep representations of dynamics (fMRI) may be as discriminative and informative as their structural counterparts (sMRI). This automatic extraction of features with minimal guidance may greatly facilitate discovering actionable causal knowledge about
the disorder by leveraging robust model introspection techniques. Model introspection, also
called model interpretation, extracts the rationale behind a model’s prediction via post hoc
feature attribution. This feature attribution results in a saliency map for every prediction
and represents the spatiotemporal activity indicative of the disorder. However, we need to
carefully design a model architecture because not every DL model is simultaneously predictive
and interpretable for time series data capturing dynamics [564].

The predictive performance of a DL model is strongly proportional to the size of training
data [134, 565], which in most neuroimaging studies is scarce to come by due to the costly
data collection process. In such a scenario, transfer learning can be a convenient approach
to dealing with this problem, as reported in numerous studies [30–34]. Although transfer
learning usually involves supervised pretraining of a model on a related task, it is difficult to
find a way to formulate the pretraining task and also the data to use so as to benefit the
downstream fMRI tasks. Model interpretation may be challenging for overparameterized
models, but if the architecture supports robust and stable sensitivity analyses [43, 44], the
interpretations for individual predictions will also be stable and robust.

The main idea of this paper is that DL can learn directly from high-dimensional signal
dynamics even in small datasets and, upon introspection, can help discover disease-specific
salient data regions, which, if carefully utilized, can advance our understanding of brain
function. To achieve this, we introduce a model that learns from dynamical data and lends
itself to interpretations. To maximally benefit from small data, we propose a self-supervised
pretraining scheme [32, 33], which maximises “mutual information local to (whole) context”
whole MILC, to capture potentially valuable knowledge from the data not directly related to the study. We use the keyword "whole" to emphasize that the self-supervised pretraining relies on the idea of mutual information maximization between the whole sequence (context embedding) and local windows (local embedding) from the same sequence. Our pretraining leverages publicly available healthy control subjects from the Human Connectome Project (HCP) [536] to establish prior knowledge about the general signal dynamics and directly transfer the insights into the downstream small data studies of schizophrenia, autism, and Alzheimer’s disease with subject age-range significantly broader than in HCP. We also propose a "Retain And Retrain" (RAR) method to validate that the biomarkers identified as explanations behind the model’s predictions capture the essence of the disorder-specific brain dynamics. A visual depiction [566] of the proposed framework is shown in Figure 6.3.

5.2 Results

We first describe all the datasets and present the results under two broad sections—whole MILC Performance and Post hoc Explanation & RAR Evaluation on FNC. The whole MILC performance indicates its predictive capacity in discriminating patients from healthy controls for each disorder separately. Post hoc explanations are feature attributions as determined by the whole MILC model for its predictions which we subsequently evaluated using the RAR scheme via an independent SVM model.
Figure 5.1: An overview of our approach to model interpretation (created in program Inkscape 1.1.2, inkscape.org/release/inkscape-1.1.2). A: Construct a model for disorder-specific discovery: we divided the entire ICA time courses into multiple sliding windows. Then we fed them into the whole MILC model that learns directly from the disorder signal dynamics and retains interpretations for further introspection. B: Leverage self-supervised pretraining to distinguish healthy subjects: learned representations assist the model in maintaining its predictive power when downstream training data is limited. C: Construct a downstream model to discriminate patients from controls for each disorder starting with the pre-trained whole MILC weights: transfer of representations learned during pretraining simplifies convergence and balances overfitting. D: Introspection of the trained downstream models: we compute saliency maps as a rationale used by the model behind every prediction using interpretability methods to extract meaningful, distinctive parts of the data. Subsequently, the estimated salient aspects of the dynamics go through an automatic validation process. To this end, we use the most salient features to retrain an independent SML model that confirms the salience of the features. This information can then be relayed to a human expert in the relevant field to interpret further and advance knowledge about the disorders. E: Examples of saliency maps as deemed highly predictive by the models for their predictions in three different discriminative tasks. Please note that the red boxes mark the highly discriminative salient parts of the data.
5.2.1 Datasets

We used the Autism Brain Imaging Data Exchange (ABIDE) [567] (569 subjects- 255 healthy controls (HC) and 314 patients) for autism spectrum disorder (ASD), the Function Biomedical Informatics Research Network (FBIRN) [550] (311 subjects- 151 healthy controls and 160 patients) for schizophrenia (SZ), and the Open Access Series of Imaging Studies (OASIS) [568] (372 subjects- 186 healthy controls and 186 patients) for Alzheimer’s disease (AZ).

5.2.2 whole MILC Performance

We evaluated the effectiveness of the proposed DL architecture with (w/) and without (w/o) the proposed self-supervised pretraining scheme, aka whole MILC, by comparing its performance against standard machine learning models. We also progressively increased the downstream sample size to investigate its impact on the model’s discriminative capacity. We used a K-fold cross-validation strategy for all the experiments below. The model was trained on samples progressively selected from the train folds, and we report the performance (AUC) on the test fold.

5.2.3 whole MILC Evaluation

Autism (ABIDE) Results (with K = 5) (see Figure 5.2 Autism spectrum panel) show that when we used a small number of subjects for training (e.g., 15 subjects per class), the pretraining improved the model’s performance compared to when the model learned only from the downstream training data ("w/o pretraining"). However, as we gradually increased the training samples, the model w/ and w/o pretraining performs almost equally. The statistical
significance test results as shown in Figure 2 further justify our observations. The reduced effects of pretraining on autism disorder classification are reasonable because the subjects from the HCP dataset are from different age groups than those from the ABIDE dataset.

**Schizophrenia (FBIRN)** Results (with $K = 5$) (see Figure 5.2 Schizophrenia panel) show that the proposed architecture w/ pretraining outperformed w/o pretraining at almost all sample sizes, and the difference was more pronounced at smaller sample sizes.

Figure 5.2: The main results from the whole MILC architecture and its comparison with standard machine learning models (SML). Apparently, the whole MILC model, in general, can learn from the raw data where traditional SML models fail to maintain their predictive capacity. Moreover, the whole MILC w/ pretraining substantially improves the latent representations as reflected in the improved accuracy compared to the whole MILC w/o pretraining. Specifically, in most small data cases, the whole MILC w/ pretraining outperformed the whole MILC w/o pretraining across the datasets. However, as expected, when we gradually increased the number of subjects during training, the effect of pretraining on the classification performance diminished, and both configurations of whole MILC did equally well. We verified this trend over three datasets that correspond to autism spectrum disorder, schizophrenia, and Alzheimer’s disease. Please note the Wilcoxon rank test results between w/ and w/o pretraining performance of the model as marked by asterisk(*) and "ns" (not significant), where $ns : p > 5e^{-2}$, $*: 1e^{-2} < p \leq 5e^{-2}$, $**: 1e^{-3} < p \leq 1e^{-2}$, $*** : 1e^{-4} < p \leq 1e^{-3}$, $**** : p \leq 1e^{-4}$.

**Alzheimer’s disease (OASIS)** Similar to what has been observed in the case of SZ (FBIRN),
the effect of pretraining on the downstream classification task ($K = 6$, to keep the testing size similar to Schizophrenia) (see Figure 5.2 Alzheimer’s disease panel) was more pronounced (comfortably outperforming) than w/o pretraining. This margin was substantial when the training data size was limited. However, as we increased the training data size, the gap between "w/ pretraining" and "w/o pretraining" was hardly conceivable.

5.2.4 Post hoc Explanation and RAR Evaluation

Once the whole MILC model was trained, we computed the feature attributions (saliency maps) as determined by the model for each prediction using model introspection techniques. These feature attribution values were estimated for every subject from the dataset because the subsequent validation depends on training and test samples. We used the RAR technique and an independent SVM classifier to validate the high discriminative power of the salient parts of data as identified by the model. Before RAR evaluation, we computed the average importance values of the overlapped time steps to obtain a single attribution value for every spatiotemporal dimension in the input sample. Refer to Figure 6.3 for example introspection maps (saliency maps) of patients from all the relevant disorder datasets.

5.2.4.1 RAR Evaluation

For RAR evaluation, we trained an SVM model on FNC matrices measured as Pearson’s correlations between time courses of the components obtained by spatial independent component analysis (ICA)[542] (discussed in Methods section). We estimated this FNC based on only 5% salient or random (baseline) data. The RAR validation results of different models trained on
three datasets with the most salient 5% (see Supplementary Fig. 1 for results from different percentages of salient data) training data are reported in Figure 5.3. As we can see, the dynamics learned by the whole MILC model were essential to maintain its predictive capacity. We observed that the model-specified salient data parts were more predictive than a similar amount of randomly chosen input data when we evaluated them for the same classification task using an independent SVM. This encouraging performance based on the salient data implies that the model can capture spatiotemporally meaningful markers suitable for patient-control distinction. Moreover, in many cases, the biomarkers identified with the "w/ pretraining" variant of the whole MILC model were more discriminative than the biomarkers specified with the "w/o pretraining" version, as reflected in the SVM's classification performance. This encouraging result generalized across the datasets, even when we used very few subjects (15) for training.

As demonstrated in classification performance shown in Figure 5.2 and validation of feature attributions shown in Figure 5.3, it is evident that the three predictive tasks were successful using our transfer learning model. In addition to quantitative validation of the automatic model introspection, we further analyzed the group-level functional network connectivity based on the model-identified salient parts of data. Refer to the connectograms (see Figure 5.4) showing the top 10% FNC computed using the most 5% discriminative data as localized by the trained model for the patients in three different disorders. We can see some interesting differences in the connectograms. Autism spectrum disorder (ABIDE) shows the least between-domain FNC highlighting within domain changes in specific cerebellum,
RAR employs SVM to classify the FNCs of the top 5% of the salient input data as estimated by the whole MILC model’s predictions. We used integrated gradients (IG) and smoothgrad integrated gradients (SGIG) to compute feature attributions. It is evident that when an independent classifier (SVM) learned on every subject’s most salient 5% data, the predictive power was significantly higher compared to the same SVM model trained on the randomly chosen same amount of data. In other words, the poor performance with randomly selected data parts indicates that other parts of the data were not exclusively discriminative as the whole MILC estimated salient 5% data parts. We also notice that sample masks over a different percentage of data coverage gradually obscured the localization of the discriminative activity within the data. Though the SVM model gradually became predictive with increased randomly selected data coverage, which we show in Supplementary Information, this performance upgrade was due to the gradual improvement in functional connectivity estimation and not attributable to the disease-specific localized parts within the data. For every disorder (Autism spectrum disorder, Schizophrenia, and Alzheimer’s disease), the higher AUC at this 5% indicates stronger relevance of the salient data parts to the underlying disorders. Furthermore, the RAR results reflect that in most cases, when whole MILC was trained with limited data, the w/ pretraining models estimated feature attributions more accurately than the models w/o pretraining.
sensorimotor, and subcortical domains [569]. Schizophrenia (FBIRN) has the most widespread predictive pattern, consistent with prior work [570] showing cerebellum interaction across multiple domains and sensorimotor changes. Finally, the predictive features for Alzheimer’s disease (OASIS) are mainly engaging visual and cognitive interactions [571]. Figure 5.5 shows full FNC matrices (based on 5% data), their disorder pairwise difference, and static FNC matrices (based on 100% data) for all disorders. As we can observe, the proposed model could capture the essential dynamics as generally captured in traditional full data FNC matrices and thus fully consistent with the knowledge from existing literature. The pairwise difference matrices imply that the different brain dynamics are indeed different for different disorders.

Furthermore, we also investigated the temporal characteristics of the saliency maps for patients and controls of each disorder. For this, we first determined the most important time points for each saliency map, expressed as temporal density and computed as the number of components for each time point that appeared in the top 5% values of the map. We observed interesting differences between groups in temporal behavior. In particular, we noticed that the temporal behavior of the most discriminative time steps is much more focused for schizophrenia and Alzheimer’s patients than their healthy controls counterparts. Put another way, the temporal density of schizophrenia and Alzheimer’s patients is generally spiky, whereas, for the healthy controls it is largely flatter. However, for autism spectrum disorder, the temporal density behavior between patients and controls is largely uniform, and the distinction, if any, is hardly noticeable. Refer to Figure 5.6 panel A for some samples showing temporal behavior of patients and controls for all disorders. To quantify these
temporal characteristics (spikiness and uniformity in temporal densities), we calculated the earth mover’s distance (EMD) [572]—a distance measure between two densities—between the temporal density computed from each saliency map and a uniform density function. The intuition behind this spread measure is that a small EMD indicates that the distribution is predominantly uniform and not localized in time, implying that the discriminatory activity is usually not confined to any specific time interval. On the other side, a large EMD indicates spikiness of the temporal behavior signaling that the discriminative activity is more focused in a shorter time interval. Refer to Figure 5.6 panel B for the distributions of EMD and corresponding statistical test results for all the disorders. We observe that the discriminative activity for schizophrenia patients is predominantly local and hence more focused in time, whereas the distinguishing characteristics of healthy controls are spread across time. We observed similar characteristics for Alzheimer’s patients. However, for autism spectrum disorder, we noticed that the temporal characteristics for both patients and controls are generally spread across time and not distinguishable. We verified our observations through a non-parametric statistical test conducted on EMD distributions for each disorder.
Figure 5.4: Top 10% FNC for patients computed using most 5% of the salient data as thresholded using feature attribution maps (saliency maps) for different disorders (created in programs MNE 1.1.dev0, https://mne.tools/dev/ and Inkscape 1.1.2, inkscape.org/release/inkscape-1.1.2). Apart from the high predictive capacity of the salient data, we observed some intriguing differences among these connectograms. The autism spectrum disorder exhibits the lowest between-domain FNC. However, salient data in autism disorder highlights domain changes in specific cerebellum, sensorimotor, and subcortical domains. The model-identified salient data reflects the most widespread pattern for schizophrenia and is consistent with the literature showing cerebellum interaction across multiple domains and sensorimotor changes. The predictive features for Alzheimer’s disease mainly concentrate on visual and cognitive interactions.
Figure 5.5: **A:** Full FNC for patients computed using most 5% of the salient data selected based on feature attribution values for different disorders. **B:** Static FNC (i.e., using 100% data) matrices for patients of different disorders. The FNC based on 5% salient data (A) does indeed convey the same focused dynamic information as currently assessed in FNC matrices based on 100% data (B). It is thus apparent that the proposed model can capture the focused information aligned with the current domain knowledge. **C:** Pairwise difference of FNC matrices based on 5% salient data. The difference FNC matrices based on focused data indicate that each disorder has a uniquely distinguishable association with brain dynamics.
Figure 5.6: A: Examples of the temporal density based on the top 5% values of the saliency maps from patients and controls for each disorder. It is noticeable that the temporal density for schizophrenia and Alzheimer’s patients is more focal in time as reflected in the spikiness, indicating that the discriminative activity for patients occurs predominantly in a shorter time interval. In contrast, for controls, model predictions do not relate to specific time intervals. For autism spectrum disorder, however, the whole MILC model did not capture any temporal adherence to the discriminative activity for patients. That is, the discriminatory events are not focal on shorter time intervals for ASD. B: The EMD (Earth Mover’s Distance) distributions as a proxy measure for uniformity/spikiness of temporal densities (edited in program Inkscape 0.92.2, inkscape.org/release/0.92.2/). We analyzed the EMD measures of patients and controls to investigate the discriminative properties of salient data in terms of the spikiness or uniformity of the temporal densities. The larger EMD measures for schizophrenia and Alzheimer’s patients substantiate that the model found the discriminative activity in shorter focused time intervals. In contrast, for ASD, the equal EMD values for both patients and controls indicate that the temporal density measures do not relate to the discriminative activity for this disorder. We verified these observations with the statistical significance (Wilcoxon rank) test results as marked by asterisk(*) and "ns" (not significant), where ns : $p > 5e^{-2}$, **** : $p \leq 1e^{-4}$.
5.3 Discussion

Standard machine learning models are widely used in neuroimaging research partly due to their familiarity and ease of use and the perceived simplicity of interpretability of the outcomes. However, this ease/simplicity takes a hit when the complexity and dimensionality of the input data are high, as is often the case with fMRI data. Our experiments (Figure 5.2) show that SML models fail to achieve good predictive performance, let alone provide meaningful interpretations of the underlying dynamics. This failure is not surprising since these proxy features are sensitive to strict assumptions about the signal dynamics [561, 562], which may only be partially accurate or accurate just under certain conditions. However, deep learning models can overcome this curse of dimensionality and learn meaningful interpretations in addition to showing high predictive performance [24, 25, 563]. This work demonstrates that DL models can achieve a deeper understanding of the underlying subject-specific signal dynamics in an fMRI setting despite the commonly expected difficulty of interpretability.

While recent advances in deep learning have proved its impressive ability to learn from a signal close to the raw data, different network architectures have benefits and limitations. The default choice of deep learning architecture for time-series data is the well-known recurrent neural network (RNN) class of models, specifically Long short-term memory (LSTM) [573]. Although LSTM models return good performance, they still have issues with interpretability due to vanishing saliency, making them unsuitable for studying multivariate signal dynamics. This necessitates building a suitable architecture that can resolve the vanishing saliency problem in the recurrent model while preserving the stability and making attributions
meaningful. To that end, Ismail, Gunady, Bravo and Feizi (2020) [564] reported that several recurrent architectures failed to provide useful attributions for the time-series data. They further reported that some architectures could extract meaningful time steps but fail to identify noteworthy features within those time steps. In this regard, we also investigated a combined CNN-RNN model and achieved high predictive performance. However, we did not find the model interpretable for time-series data. Instead, we found multi-level hierarchical attention on top of LSTM as used in whole MILC useful for interpretable time-series prediction. Results show that our \textit{whole} MILC model resolves the vanishing saliency problem and is a good tool for introspection of the multivariate signal dynamics.

Interpretation of deep learning models may uncover domain-specific knowledge [37, 38] that would otherwise require high cost, effort, and time investments. Often, it may also assist in identifying if the model has inherited any inherent bias from the data. On the other hand, some studies [40, 41] raised doubts about the transparency of deep learning models and the applicability of popular interpretability methods. Notwithstanding these diverging opinions, the significance of interpretability and visualization in medicine and healthcare cannot be overstated [150] and should involve medical experts as well. Expert human involvement in interpreting the extracted information on clinical terms may help validate and guide disease-associated discovery. A recent review [574] reveals that deep learning models are a viable clinical supportive tool in the neuroimaging domain. However, studies have concentrated mainly on structural imaging data. Conversely, this paper introspects deep learning models for multivariate time-series data, which we think is an essential step toward interpretability.
research of functional imaging data. To this end, our model introspection results reveal the capacity of the proposed model to locate highly predictive disease-relevant information. Specifically, we validate the efficacy of the estimated feature attributions by proposing a method called RAR. With RAR and an independent SML model, we verify that IG and SGIG, when applied to whole MILC model, are robust, stable, and can demonstrably identify disorder-relevant parts of the brain dynamics. Precisely, the model-identified features offer very high predictive performance compared to random baselines for schizophrenia, Alzheimer’s disease, and autism spectrum disorders. Moreover, our FNC analysis on model introspection results, as shown in Figure 5.5, harmonizes with the prior work [569–571] for all the disorders.

We analyzed the required "what" and "when" aspects of the discriminative dynamics the model captured for patient-control distinction. Toward this goal, FNC analysis on the salient data revealed the minimally required connectivity ("what") of the discriminative dynamics that the model used to distinguish patients from controls. We further investigated if the model leveraged any temporal ("when") information for its discriminating power. Accordingly, we analyzed when, if such information exists, the discriminative events happen and how this temporal behavior changes between patients and controls for each disorder. As such, we analyzed the temporal densities computed from salient 5% data. Interestingly, for schizophrenia and Alzheimer’s disorders, we observed that the model used temporally dense information to distinguish patients from controls. However, no temporal association is noticed in the model behavior to distinguish ASD patients from controls. We substantiate this aspect of temporal association using a non-parametric statistical test as shown in Figure
Deep learning models typically require large amounts of data for efficient training. However, in the field of neuroimaging, collecting massive amounts of homogeneous data is infeasible thus constraining researchers to work with small data. In such cases, transfer learning [30–33] is practically helpful to enable learning directly from data. Self-supervised learning has made significant progress in computer vision classification tasks [34] and is equally applicable to deep convolutional and recurrent networks. As demonstrated, our self-supervised pretraining scheme [32] enables downstream learning with minimal training data, making the direct investigation of system dynamics feasible. Our findings demonstrate that self-supervised pretraining on healthy adults dataset noticeably uplifts the downstream model’s performance on a disparate disorder dataset. These benefits generalize across datasets and disorders and thus alleviate the need to collect a massive amount of expensive data.

While the proposed framework is a stepping stone toward the direct study of signals, the proposed approach still needs to be improved to make it a clinically relevant. An interpretable model is essential to grasp better the difficult task of interpreting brain dynamics of mental disorders, and our approach demonstrably works quite well and provides a promising utility. However, a possible drawback of this current work is that the classification performance in some cases may be suboptimal due to learning directly from the signals with minimal guidance. Moreover, the spatial maps have been left unexplored. That is, utilizing only the time courses could slightly bias our models to pay more attention to the temporal component of the signal. In the future, we would like to scale our models to be able to handle full brain
raw fMRI data without ICA pre-processing. We hope our interpretability approach will become even more informative in that case.

To conclude, we interpret DL models trained on fMRI signals to classify mental disorders from healthy controls to provide means to identify salient parts of the brain dynamics (activity patterns). In particular, we show that one can capture the dynamic signatures as generally captured in traditional full data functional network connectivity (FNC). We further demonstrate that the brain function manifests itself via unique dynamic signatures across time scales (latent temporality) in various disorders. Subsequently, we present an adaptive, interpretable methodology to capture these temporally transient dynamic signatures that can help distinguish disorders. Understanding the spatial and temporal specificity of the brain activity patterns will help establish the technique for clinical use by relating the differences in signature to symptoms. Moreover, to achieve these desirable disorder-specific insights, the proposed pretraining method waives the need for well-defined ground truth (biomarkers) about the disorder under consideration and a larger sample size. In the future, this method could be a significant step towards establishing more robust correlates of function-structure dependency in the brain and can also be applied more broadly to understand inter-and intraindividual variability and alterations across psychiatric disorders.

5.4 Methods

The proposed methodology consists of 4 steps: model pretraining, downstream classification, feature importance estimation, and feature evaluation. First, we pre-trained the proposed
network (whole MILC) [32] on a large unrelated and unlabeled dataset to learn valuable latent representations. This pretraining, as described in the whole MILC Section, intuitively lets the network learn foundational knowledge about the dynamics only from the healthy subjects. For pretraining and downstream tasks, we used the same model as used in [32]. However, for the current study, we replaced the CNN encoder with a recurrent encoder because we found it more stable for post hoc explanations of multivariate time-series data while interpreting the model’s predictions. As the learned dynamics are directly transferable, we used the pre-trained network to discriminate patients from healthy controls in different downstream tasks. In the second step, we trained the downstream classification model to learn more from the downstream training data dynamics. In the third step, we estimated feature importance values based on the model’s predictions using different interpretability methods (see Model Interpretability section). In the fourth step, we evaluated the estimated features using RAR method and an SVM model as described in the RAR Section. Before going through the methodological pipeline, we preprocessed the data as described below.

We state that the study was performed according to all relevant guidelines and regulations. While the original data were collected under approved IRB protocols by the original study teams, we were not involved in this step. The data were provided to us as anonymous. We submitted the proposed work to the GSU IRB which designated the project as ‘not human subjects’ thus there was no need for ongoing IRB oversight of the project.
5.4.1 Preprocessing

We preprocessed the raw resting-state fMRI data using statistical parametric mapping (SPM12, http://www.fil.ion.ucl.ac.uk/spm/) package in MATLAB 2016 environment. We removed the first five scans for the signal equilibrium and participants’ adaptation to the scanner’s noise. We performed rigid body motion correction using the toolbox in SPM to correct subject head motion, followed by the slice-timing correction to account for timing difference in slice acquisition. Using an echo-planar imaging (EPI) template, the fMRI data were subsequently warped into the standard Montreal Neurological Institute (MNI) space using an echo-planar imaging (EPI) template. We resampled the fMRI data to $3 \times 3 \times 3$ mm$^3$ isotropic voxels and further smoothed using a Gaussian kernel with a full width at half maximum (FWHM) = 6 mm. After the preprocessing, we selected subjects with head motions $\leq 3^\circ$ and $\leq 3$ mm in the analysis. To ensure high data quality, we performed quality control (QC) on the spatial normalization output and removed subjects with limited brain coverage [575]. We used ICA time courses as these offer a better representation of the data than anatomical or fixed atlas-based approaches [576]. For each dataset, we used ICA components derived via a fully automated approach [553]. In this framework, we performed spatial group ICA on two independent datasets with a large sample of healthy controls (human connectome project [HCP, 823 subjects after the subject selection] and genomics superstruct project [GSP, 1005 subjects after the subject selection]) to generate network templates. For each dataset, we conducted group ICA, respectively. The estimated ICs from the two datasets were then matched by comparing their corresponding group-level spatial maps. Those pairs
are considered consistent and reproducible across datasets if their spatial correlation is $\geq 0.4$. We characterized a subset of these matched ICs as ICNs instead of physiological, movement-related, or imaging artifacts. Five fMRI experts carefully examined those matched ICs, and ICs with more than three votes were considered meaningful ICNs. The experts evaluated the ICs based on the expectations that ICNs should have their activation peaks in gray matter and low spatial overlap with known vascular, ventricular, motion, and other artifacts. ICNs also should have dominant low-frequency fluctuations on their corresponding time courses (TCs). We used these meaningful ICNs as network templates for further individual-level ICA analysis. We obtained 100 ICA components for each dataset using the same procedure as described in [575]. However, this study used 53 intrinsic networks (components) for all experiments because they perfectly matched the standard network templates. In pretraining, we used a sliding window of $53 \times 20$ size with stride $= 10$ along the time dimension to feed the ICA time courses through a parameter-shared encoder. In all downstream classification experiments, we used a similar sliding window with stride $= 1$.

5.4.2 Whole MILC

The *whole* MILC model, as shown in Figure 5.7, consists of two unidirectional LSTM models arranged in a top-down fashion. While the low-level LSTM functioned as a parameter-shared encoder for the sliding window over ICA time courses, the top-level LSTM used the encoder embeddings to generate a global representation for the entire sequence. Both LSTM models separately applied an attention mechanism [91] to retain interpretable information for further model introspection. One of the benefits of the *whole* MILC model is that it
is pre-trainable. Moreover, the learned representations are directly transferable to a set of downstream discriminative tasks. The whole MILC model used a self-supervised pretraining objective [32] that maximized the mutual information between the latent space of a window (time slice from ICA time courses) and the corresponding whole sequence (complete ICA time courses per subject).

Let \( D = \{ (u_i^t, v^j) : 1 \leq t \leq T, 1 \leq i, j \leq N \} \) be a dataset of window-sequence embedding pairs computed from ICA time courses, where subscript \( t \) refers to the \( t \)-th window, superscripts \( i, j \) each refers to a sequence number. \( T \) is the number of windows in a sequence, and \( N \) is the total number of sequences in the dataset. \( D \) can be decomposed into a set of positive pairs \( D^+ \) (\( i = j \)) and a set of negative pairs \( D^- \) (\( i \neq j \)) denoting a joint and a marginal distribution respectively for the window-sequence pairs in the latent space. With a separable function \( f \), we used InfoNCE estimator [577] to compute a lower bound \( I_f(D^+) \) on the mutual information defined as:

\[
I_f(D^+) \geq \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \log \frac{\exp f((u_i^t, v^j))}{\sum_{k=1}^{N} \exp f((u_i^t, v^k))}}{N}, \tag{5.1}
\]

\( f \) was defined as \( f(u, v) = \phi(u_i^t)^T(v^j) \), where \( \phi \) was some embedding function learnt by network parameters. \( f \) learned an embedding function such that it assigned higher values for positive pairs than for negative pairs, i.e., \( f(D^+) \gg f(D^-) \). To make it precise, \( u_i^t \) and \( v^j \) in the Equation 5.1 respectively refer to window embedding \( z_i^t \) and global sequence embedding \( c \) in Figure 5.7. The InfoNCE loss using \( f \) as a representation model is defined as \( L = -I_f \).
Figure 5.7: The whole MILC architecture—an attention-based top-down recurrent network (created in programs Adobe Illustrator 26.0.3 and Inkscape 1.1.2). Precisely, we used an LSTM network with an attention mechanism as a parameter-shared encoder to generate the latent embeddings $z$ for the sliding window at all relevant positions. The top LSTM network (marked as LSTM) used these embeddings ($z$) to obtain the global representation $c$ for the entire subject. During pretraining, we intended to maximize the mutual information between $z$ and $c$. In the downstream classification task, we used the global representation $c$ directly as input to a fully connected network for predictions. Based on these predictions, we estimated feature attributions using different interpretability methods. Finally, we evaluated the feature attributions using the RAR method and an SVM model.
5.4.3 Attention Mechanism

The attention mechanism is a valuable construct commonly used in DL architecture to preserve long-term dependency in the recurrent neural network. Initially, Bahdanau, Cho, and Bengio (2014)[91] introduced the attention mechanism for the neural machine translation to compute the relevance of source words toward each output word. However, the attention mechanism can benefit other applications too. For example, we used the attention mechanism to solve vanishing saliency problems in the LSTM networks to retain interpretable information during model training. In the attention mechanism as used in whole MILC model, we took all the hidden states \( h = [h_1, h_2, \ldots, h_n] \) from the LSTM network and concatenated each hidden state \( h_i \) with the hidden state at the last time step \( h_n \) before passing through an attention mechanism \( f_a \). The attention mechanism \( f_a \), similar to the additive attention mechanism introduced in [91], took pairs of hidden states \( (h_i, h_n) \) as inputs, passed through a 2-layer feed-forward network and generated a vector of \( n \) alignment scores \( f_a(h_i, h_n) \). The alignment score for each time point \( i \) intuitively indicates the degree of relevance of the corresponding hidden state to the overall embedding. We normalized the alignment scores using softmax to produce a series of weights \( \alpha_1, \alpha_2, \ldots, \alpha_n \). \( \alpha_i \) for each time point is defined as:

\[
\alpha_i = \frac{\exp(f_a(h_i, h_n))}{\sum_{k=1}^{n} \exp(f_a(h_k, h_n))} \tag{5.2}
\]

where \( n \) was the number of time steps over which attention was applied. Note that the value of \( n \) for the encoder LSTM network (for the sliding window) differed from the top LSTM network (for the full subject). The global representation \( c \) (or the window embedding
\( z \) was generated using the formula as follows:

\[
    c = \sum_{k=1}^{n} \alpha_k h_k \tag{5.3}
\]

### 5.4.4 whole MILC Setup

**Encoder Embedding:** The LSTM encoder with an attention mechanism used a sliding window of 53 \( \times \) 20 size to feed the ICA time courses and encoded features at each time point into a 256-dimensional representation. At each position of the sliding window, we concatenated the hidden state for each time step \( t_i \) within the window with the final hidden state of the same window as described in the attention mechanism. We then passed these concatenated 512-dimensional vectors through an attention network, a two-layer feed-forward network with hidden units 64. The network learned a series of weights representatives of the magnitude of attention regarded as important for the time steps. All the hidden representations within a window were then weighted based on the attention scales to produce window embedding \( z \).

**Pretraining:** In *whole* MILC based pretraining, we passed all the encoder embeddings \( z = z_1, z_2, \ldots, z_n \) to another unidirectional LSTM network with an attention mechanism. In this top recurrent network, each window embedding \( z_i \) corresponded to the input for a single time step. We used 200 dimensions to represent the hidden state for this top network. We concatenated each hidden state with the hidden state at the last time step to make it contextually relevant for the attention mechanism. The top attention network used 400 input neurons and 128 hidden units to learn \( k \) weights, where \( k \) was the number of input windows.
These weights were used as coefficients in the linear combination of hidden representations to generate a global embedding $c$ of dimension 200 for each subject. Based on $c$ and $z$, we pre-trained the neural network to maximize the mutual information between a window and the corresponding input sequence. We used subjects from the HCP dataset for pretraining and used 700 subjects for training and 123 subjects for the test, obtaining 89% pretraining accuracy.

**Classification Tasks:** In downstream tasks, we deal with classifying subjects into patients and controls separately for each disorder. Similar to pretraining, we fed ICA time courses into the LSTM encoder using a sliding window. The LSTM encoder projected all the windows into latent representations $z$, which were then passed to another LSTM network to obtain a global representation $c$. Finally, on top of $c$, we used a feed-forward network with 200 hidden units to perform binary classification. We gradually increased the number of supervised training subjects to observe the pretraining effect on downstream data size compared to the setup where we used no pretraining. For each experiment, we report cross-validated results. Moreover, we performed ten repetitions of each experimental setup, with different random seeds for every cross-validation fold to ensure stable results. For each random seed, we randomly chose the training samples as required from the available training pool.

### 5.4.5 Model Interpretability

We describe an input as a vector $\mathbf{x} \in \mathbb{R}^d$. Let us define the deep learning model as a function $F : \mathbb{R}^d \rightarrow \mathbb{R}^C$, where $C$ is the number of classes in the downstream classification problem (in our case $C = 2$). Moreover, let us also assume that the mapping $F_c(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ defines
the class-specific logit, where $c$ is the predicted class. An explanation method, also called model introspection method, provides an explanation map $E : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that maps $\mathbf{x}$ to a saliency map of the same shape. Values in the saliency map correspond to the ‘relevance’ or ‘importance’ of that dimension to the model’s prediction.

The need to enable model interpretation led to a variety of model introspection techniques that can be roughly split into three groups: 1) model-sensitive [43, 44], 2) model-agnostic [49, 532], and 3) counterfactual explanations [80]. The techniques have their relative benefits and pitfalls in addressing the desiderata of different applications [578]. Adebayo, Muelly, Liccardi, and Kim (2020) [579] reported that, under normal conditions, gradients, smoothgrad [44], and integrated gradients (IG) [43] passed end-user recommendations. Additionally, the smoothgrad method [44] resolves the problems [580] of saliency maps, which in general, are susceptible to noise and input perturbations. Guided by these findings, we relied on IG, and smoothgrad on IG to introspect the proposed model. Notably, we found IG and smoothgrad on IG generalizable, stable, and noise-robust across the disorders.

5.4.6 Random Baseline

We randomly assigned feature importance values to create random baselines to validate the post hoc explanations (saliency maps). Specifically, we ordered the features uniformly at random using random permutations and considered each permutation as an order of importance. We refer to this random estimator as $g^R$ throughout the paper. In contrast, we used the magnitude of the estimated attribution values as the order of importance for the model-generated post hoc explanations. To evaluate the efficacy of the estimated feature
importance, we compared the predictive power of the model-estimated salient features against random baselines using a technique called RAR, which we describe below.

5.4.7 RAR Method and Setup

In RAR, we retained only a small percentage of the most salient features as determined by the model and replaced other features with non-informative values (zeros). We used these modified samples to retrain an SVM model to evaluate the effectiveness of the estimated feature attributions. In particular, we show that the performance obtained with whole MILC model-estimated salient features far exceeded the random baseline. We mathematically describe the RAR scheme as follows:

Let us define $X$ to be the original dataset. $X^M|g^R$ be the modified dataset based on random importance estimates and $X^M|g_i$ be the modified dataset according to the saliency maps generated by applying some interpretability method $g_i$ on whole MILC predictions. We computed static functional network connectivity, measured as Pearson’s correlation coefficients, for each sample in $X^M$. We used these correlation coefficients as features to train an independent SVM model de novo. We evaluated the classification performance of the SVM models trained separately with whole MILC-generated salient features and randomly selected features. Indeed, we show that $\xi(X^M|g_i) > \xi(X^M|g^R)$, where $\xi$ is the performance evaluation function, e.g. area under the ROC curve and/or accuracy.

It is to note that we sorted the features based on their signed attribution values before considering them for validation. We searched for the SVM (nonlinear) parameters using a parameter grid and 3-fold cross-validation on the training data. We used the same folds and
train-test splits for the RAR evaluation as used in the *whole* MILC model. Figure 5.8 shows the schematic of the end-to-end process: 1) training the *whole* MILC and feature attributions and 2) Evaluation of the feature attributions using RAR and an SVM model.

\[
\xi(X^M \mid g_i) > \xi(X^M \mid g^R),
\]
where \(\xi\) is the performance evaluation function (e.g., area under the curve) and \(X^M\) refers to the modified dataset constructed based on only retained feature values.

5.5 Conclusion

In this chapter, we discuss our interpretable deep learning framework suitable for learning dynamics directly from data. The deep learning model is pre-trainable and can take advantage of unlabeled and unrelated data to help neuroimaging studies suffering data scarcity. We pretrained our DL and applied them to three downstream mental disorders—schizophrenia,
autism, and Alzheimer’s disease. The pretrained model achieved improved performance in these downstream tasks than its “w/o pretraining” counterparts. The framework supports interpretability and is able to extract disorder-relevant dynamics of the brain disorder. We also discussed our proposed RAR validation framework that can quantitatively validate the predictability of the generated explanations. We also discussed our analysis of the recognizable temporal behavior of the salient regions of the patients and controls. The preliminary insights from the analysis suggest a new avenue to advance our understanding of disorder dynamics.
Interpretability methods for deep neural networks mainly focus on the sensitivity of the class score with respect to the original or perturbed input, usually measured using actual or modified gradients. Some methods also use a model-agnostic approach to understanding the rationale behind every prediction. In this paper, we argue and demonstrate that local geometry of the model parameter space relative to the input can also be beneficial for improved post-hoc explanations. To achieve this goal, we introduce an interpretability method called ‘geometrically-guided integrated gradients’ that builds on top of the gradient calculation along a linear path as traditionally used in integrated gradient [43] methods. However, instead of integrating gradient information, our method explores the model’s dynamic behavior from multiple scaled versions of the input and captures the best possible attribution for each input. We demonstrate through extensive experiments that the proposed approach outperforms vanilla [52, 53] and integrated gradients in subjective and quantitative assessment. We also propose a "model perturbation" sanity check to complement the traditionally used "model randomization" test.

6.1 Introduction

The past decade has seen a wealth of new advancements in deep learning (DL), improving performance in a wide array of possible problems in many areas, especially in classification tasks from computer vision and natural language processing [581, 582]. However, this improved performance comes with a cost, the models are not easily explained or described
with simple observations. This lack of intelligibility hinders wide-scale use of these models in safety-critical domains such as healthcare, education, the justice system, and many others. This need for decipherability has given way to parallel advancements in a subfield of machine learning known as interpretability. Interpretability has grown as quickly as the DL field itself, providing many new state-of-the-art techniques [43, 50, 52–55, 58, 82, 90].

Interpretability research generally tackles these problems by either describing the model’s decision-making processes or by generating post-hoc explanations, either for the model as a whole, or for each sample. There are many post-hoc explanation methods from research spanning the last 2 decades. The performance and properties of these methods vary widely across different architectures and domains. In recent years, beginning with integrated gradients (IG) [43], some of these methods [58, 90, 583] have satisfied two important properties: sensitivity and implementation invariance.

These post-hoc analysis methods, specifically the ones that leverage model gradients to explain each sample, are known as saliency techniques. Vanilla gradients (GRAD) [52, 53] and IG, two of the most popular methods, can be noisy [44, 82, 83, 114]. We show an example of this noise in Figure 6.1. GRAD specifically violates both implementation invariance and sensitivity. IG also has specific noise relating to the averaging or integrating over the interpolation path. Other sources include the large curvature of the network’s decision function [584], numerical approximation of integration [90], and baseline choices [114, 583]. In order to both conform to the two aforementioned properties and reduce noise, we suggest that the gradients can be improved by leveraging more aspects of the model space than
the exact sample loss or a linear interpolation of the sample as is found in IG. Specifically, geometric properties of the loss landscape, which have previously been utilized to strengthen model performance and robustness [585–588].

We propose *Geometrically Guided Integrated Gradients (GGIG)*, an algorithm that builds on top of traditional IG to reduce noise. We expand IG by ascending through the loss space by maximizing the class-specific logit, much the way class activation maximization [589] works. This ascension, or maximization of the class logit, allows GGIG to find the gradients that are most discriminative, which we suggest is a valuable property for any interpretability mechanism. After describing the method, we empirically show that GGIG improves the quality and robustness of the saliency maps. Figure 6.2 describes the general pipeline of the work.

![Figure 6.1: Comparison of gradients, integrated gradients and GGIG based on top 5% pixel attributions for the "gazelle" prediction. In GGIG explanation, the details of the animal are clearly noticeable and discarded the undesirable regions from the explanation.](image)

Our main contributions are as follows:

- We propose GGIG an interpretability method that starts with a linear path as used in IG and finds the path that enhances the class activation for the underlying prediction.
• We propose a model perturbation sanity check also called "\( \sigma \)-perturbation" that we think all explanation methods should satisfy.

• We also propose another sanity check called the "target object sensitivity" test, where we require interpretability methods to exhibit sensitivity only to the target objects, not to the background or other irrelevant parts.

• We show that the proposed method offers better saliency maps for different datasets and architectures when assessed through visual inspection and quantitative metrics.

6.2 Related Work

The obfuscatory nature of DL models is well documented and has been a popular research problem for over a decade. Many studies have proposed solutions with varying quality, costs, and benefits [43, 44, 48, 50–55, 82, 86, 583, 590]. Gradient-based methods, also referred to as visualization methods [57], are easy to implement and applicable to all models that use gradient descent.

GradCAM [50] identifies the focal regions pretty well, and maps are highly predictive. However, the saliency maps are blobby [58] and usually expand around the actual objects. While IG has its own specific problems, several recent studies have refined IG attributions because this method has many desirable properties. Kapishnikov et al. [58] proposed a region-based attribution method, called XRAI, that mainly refines the IG attribution based on attribution density. However, XRAI requires a way to cluster the input features, which may not be available for different data modalities. Kapishnikov et al. [90] proposed another
method, called GIG, which provides an adaptive path method based on input, baseline, and the model. This method starts at the baseline and selects only those pixels with the lowest partial derivatives to take closer to the next interpolation point, thus avoiding the gradient accumulation from saturation regions. In other words, it constitutes the path based on dynamic projections of the linear interpolation path.

To reduce the inherent noise in the saliency maps, we may utilize some useful loss landscape properties as observed in several studies [585–588] to design reliable interpretability methods.

Figure 6.2: Overall pipeline of the work. First, we generate explanations for each prediction. Secondly, for quantitative evaluation, we start with a complete defocused image and combine with only salient pixels (thresholded at $x\%$) to create a test image. Finally, we feed the model with the test image and measure both accuracy and relative softmax scores.
Algorithm 1: Geometrically Guided Integrated Gradients

**Input:** $F$ = model logit function, $x$ = sample

**Parameter:** $n$ = interpolation points, $x'$ = baseline, $m$ = ascending iteration, $lr$ = learning rate

**Output:** $e$ (GGIG explanation)

1. gradients $G ← \{\}$
2. for $k ← 0$ to $n$ do
3.   $x^0_k ← x' + \frac{k}{n}(x - x')$
4. for $j ← 0$ to $m - 1$ do
5.   $x^{j+1}_k = x^j_k + lr \times \nabla F(x^j_k)$
6.   $G \leftarrow \nabla F(x^j_k)$
7. end for
8. end for
9. $e = \max(G)$  \# pixel-wise maximum attribution
10. return $e$

6.3 Geometrically Guided Integrated Gradients

We propose GGIG a method which incorporates the idea of path methods [43] and enhances the quality of the attribution by analyzing the local loss behavior. Figure 6.3 shows the schematic diagram illustrating the functional mechanism of GGIG.

Like IG, GGIG starts from the baseline $x'$ and constitutes a linearly interpolated path. However, instead of accumulating gradients along the path, it updates the path gradients in the direction where the model maximizes class activations. GGIG thus maximizes class activation for each of the linearly interpolated points. The procedural steps for GGIG are shown in Algorithm 1. We hypothesize that the prediction curve in the vicinity of $x$ holds important information about the interaction between model $f$ and input $x$.

Let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined as the mapping from the input space to the class-specific logit and $b_0, \ldots, b_n$ be a set of $n$ linearly interpolated points for each sample between the baseline,
Figure 6.3: The working mechanism of GGIG. It takes a baseline \( x' \), creates a linearly spaced path to the actual input \( x \). From each interpolation point, it follows a trajectory in the loss landscape that maximizes class activation.

\( b_0 \) and the exact sample \( x \), where \( b_i = b_0 + \frac{i}{n}(x - b) \). For each \( b_i \), we compute a form of gradient ascent for \( m \) iterations over the given interpolated sample (as opposed to the model), defined as: \( b_{i+1}^j = b_i^j + lr \times \nabla F(b_i^j) \) where \( j \) is the incremental ascent over the sample and \( lr \) is the learning rate. This ascending mechanism allows the model to take the direction where logit values are enhanced. All gradients over the ascension, \( \nabla F(b_i^j) \), are collected, and the
max $\nabla F(b_j^i)[p]$ (for each pixel $p$), where $i = 0, 1, 2, \ldots, n$ and $j = 0, 1, 2, \ldots, m$, is selected as the final attribution for the pixel $p$.

IG [43] has many desirable properties and we have used IG as the baseline to compare the proposed method. We used slightly different formulation of IG. Generally, IG uses interpolation technique to integrate importance at different discrete intervals between uninformative baseline, say $\bar{x}$ and the input $x$, to give an integrated estimate of feature importance. IG based feature importance is computed as:

$$e = (x - \bar{x}) \times \sum_{i=1}^{k} \frac{\partial F(\bar{x} + \frac{i}{k} \times (x - \bar{x}))}{\partial x} \times \frac{1}{k}$$

(6.1)

The ultimate estimate $e$ depends on the value of $k$ (number of intervals) and the choice of a suitable uninformative baseline $\bar{x}$. The traditional integrated gradients scale raw attributions (operand on the right of the multiplication operator) by $x - x'$ (operand on the left).

**Element-wise multiplication is misleading:** Adebayo et al. [109] observed that element-wise multiplication could be misleading. This misleading happens mainly because the input dominates the product even with drastic changes in gradient vectors. So, the interpretability methods with this element-wise multiplication component in their formulation can provide input-dominant explanations that may deceive human understanding.

Furthermore, Ancona et al. [70] suggested that this point-wise multiplication was initially justified to sharpen the gradient explanations; however, it is better justified when a measure of salience is a priority over mere sensitivity. In this paper, we were more interested in the sensitivity of features rather than their marginal salience to the target score. For all of these valid reasons, we did not multiply the integrated gradient with $x - x'$ to avoid
input dominance from the explanations. Moreover, we did not consider other gradient-based methods like grad $\odot$ input for the same reason.

The baseline can be problematic for correct attributions [58, 114]. Kapishnikov et al. [58] addressed the issue of baseline by using both black $(0, 0, 0)$ and white $(1, 1, 1)$ baselines. While baseline may be an issue for traditional IG formulation, avoiding direct or modified (as in IG) element-wise multiplication resolves the baseline issue from this work.

### 6.4 Model Perturbation Sanity Check

In this section, we propose a model perturbation sanity check for attribution methods. We refer to this perturbation as $\sigma$-perturbation. This perturbation seeks two important properties:

**Property 1.** Let $M$ be a model with the parameter vector $w = [w_1, w_2, \ldots, w_n]$ and $\sigma$ be the standard deviation of $w$. Let $x$ be the sample for which we are generating explanation. Let $S_M(x)$ be the probability score $M$ generates for the input $x$. Let $M'$ be a model obtained by perturbing the model $M$ using $w' = w + \mathcal{N}(0, \epsilon I)$, where $0 \leq \epsilon \leq \sigma$. For sufficiently large perturbation level $\epsilon \gg 0$, $S_{M'}(x)$ should reduce to $1/C$, where $C$ is the number of classes.

**Property 2.** Let $e_M = E_M(x)$ be an explanation for the sample $x$ generated by the original model $M$ and $e_{M'} = E_{M'}(x)$ be an explanation for the same sample generated by the perturbed model $M'$. Also, let $S(e_1, e_2)$ be any similarity measure between two explanations $e_1$ and $e_2$. With gradual perturbation of the model $M$, the similarity between $e_M$ and $e_{M'}$ should evaporate accordingly. For sufficiently large perturbation level, i.e., $\epsilon \gg 0$, the $S(e_M, e_{M'}) \approx 0$. 
We claim that all attribution methods should satisfy this sanity check to ensure their meaningfulness and sensitivity to the model parameters. We applied this sanity check to vanilla gradients, integrated gradients and GGIG. Interestingly, all of the methods satisfied the $\sigma$-perturbation sanity check.

We report $\sigma$-perturbation sanity check results in Figure 6.4.

6.5 Experiments

In this section, we study the performance of GGIG on different benchmark datasets—MNIST and ImageNet. For MNIST, we created and trained a convolutional neural network. However, for the ImageNet dataset, we used popular pre-trained models like Inception v3 and ResNet 101. We assessed the comparative performance of GGIG against vanilla gradients, integrated gradients, random baseline, and edge detectors. We adopted two popular evaluation metrics for quantitative comparison of post hoc interpretability methods, which we discussed later.

We conducted all the experiments using 1 NVIDIA GPU per job, and each job used 2
CPUs. We assigned 40GB of memory per node running a single job. Please note that a single job refers either to a single model training or running an experiment to generate post hoc explanations using a single interpretability method on the selected subset of images for a single dataset.

6.5.1 MNIST Experiments

Training on MNIST Dataset For MNIST [591], the model architecture was a CNN that consisted of two convolutional layers with (32, 64) filters of sizes (5, 5). Each convolutional layer is followed by a $2 \times 2$ max pooling layer and a ReLU activation. We fed the final convolution output to a fully connected network with 1024 input and 10 output units (softmax). We optimized the model using stochastic gradient descent (SGD) with a learning rate of 0.0004 and momentum of 0.9. The model was trained for 400 iterations with a mini-batch size of 64 and finally achieved an accuracy of 99.2%.

Post hoc explanation experiments For GGIG, we used a learning rate of 0.0001 for gradient ascent from each linear interpolation point between input $\mathbf{x}$ and baseline $\mathbf{x}' = 0$. For MNIST, we iterated the gradient ascent for 200 steps and noted the maximum sensitivity along the gradient ascent trajectory for each input. We display the saliency maps obtained on the MNIST dataset in Figure 6.5.
Figure 6.5: **Top:** Selected maps for MNIST samples generated using different methods. **Bottom:** Quantitative evaluations using different correlation and similarity metrics. It is obvious that the maps produced by GGIG have higher structural and numerical correlations with the input. Precisely, GGIG, as an explanation method, performs the best by a large quantitative margin.
Table 6.1: Summary of Quantitative Evaluation on MNIST dataset

<table>
<thead>
<tr>
<th>Saliency Method</th>
<th>Spearman Rank Correlation</th>
<th>Weighted Jaccard Similarity</th>
<th>Structural Similarity</th>
<th>Normalized Mean Square Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median</td>
<td>std</td>
<td>median</td>
<td>std</td>
</tr>
<tr>
<td>GRAD</td>
<td>0.525</td>
<td>0.048</td>
<td>0.210</td>
<td>0.035</td>
</tr>
<tr>
<td>IG</td>
<td>0.516</td>
<td>0.049</td>
<td>0.214</td>
<td>0.035</td>
</tr>
<tr>
<td>GGIG</td>
<td><strong>0.533</strong></td>
<td>0.056</td>
<td><strong>0.332</strong></td>
<td>0.041</td>
</tr>
</tbody>
</table>

6.5.2 Quantitative Evaluation on MNIST Dataset

Visual inspection of explanation methods can be unreliable as it is possible to create adversarial samples [110, 111] that can fool the human eye, totally changing the model predictions. We perform different similarity measures between the maps and the input to understand the quality of the proposed methods on MNIST, namely, Spearman Rank Correlation, Weighted Jaccard Similarity, Structural Similarity, and Normalized (Reverse) Mean Square Error. For quantitative evaluation, both data and maps were rescaled in the range [0, 1]. We assumed that the amount of information a method can capture about the structure and distribution of the input in the saliency maps directly determines its quality as an explanation method. We report the quantitative evaluation in Table 6.1. Figure 6.5 shows sample maps and the detailed results of quantitative evaluation. As we can observe, GGIG outperforms other methods. Comparatively, GRAD and IG retain little information about the numerical and structural association to the input.
6.5.3 Experiments on ImageNet

We also evaluated the proposed methods using a subset of images from the ImageNet dataset [592] and different pretrained models, namely ResNet-101 [593], and Inception V3 [594]. Though we found meaningful maps in every case, maps still vary in quality possibly due to their architectural differences. Figure 6.6 and Figure 6.7 show some maps produced using different saliency methods and pretrained models. A clear demonstration of difference GGIG offers compared to other gradient-based methods are shown in Figure 6.8.

GGIG needs few hyperparameter choices, especially as it involves gradient ascent and linear interpolations. To preserve simplicity and reproducibility, we used zero baselines and 50 interpolation points in all experiments. As a model’s loss landscape is highly unsmooth, we used a lower learning rate of $1e^{-4}$ and iterated over 200 gradient ascent steps for all the experiments.
Figure 6.7: Comparative saliency maps for Inception V3 model and ImageNet dataset generated by GRAD, IG, GGIG, and Edge Detector. The maps obtained using GGIG are more discriminative and more clearly reveal the underlying structure of the class-associated objects.

6.5.4 Target Object and Model Parameter Sensitivity Tests

Saliency methods are not edge detectors: Adebayo et al. [109] observed that many saliency methods, including vanilla gradients and integrated gradients can appear like edge-detectors for 1-Layer Sum-Pool Conv Model. This assumption may only hold for shallow models. To analyze the edge-detector like behavior for deep models, we conducted an experiment where we replaced the original background of the images with a sharp-changing image. In particular, we assigned ImageNet samples a very different fixed background (a black and white checkerboard) and generated post-hoc explanations using GRAD, IG, and GGIG. We show the resulting explanations in Figure 6.9. It is obvious from the resulting maps that the model used concepts, not merely edges from the training objects. Moreover, as expected, all the saliency methods ignored the background and only attributed the object’s
Figure 6.8: The input image is from ImageNet: we show the input image and 10% salient pixels of the explanations (saliency maps) generated using 'GRAD', 'IG' and 'GGIG' (our method). As we can see, while GRAD and IG are noisy and not focusing on any discriminative parts of the image, GGIG is demonstrably focusing on very distinguishable regions which are representative of a gentleman. For example, salient predictive regions include facial parts, tie, collar, wrist watch etc, which is highly aligned to human understanding of the input image class.

parts for prediction explanations. In fact, we showed that the methods exhibited sensitivity to the target objects; hence, we call it the "target object sensitivity" test. It is apparent
Figure 6.9: "Target Object Sensitivity" test experiments on ImageNet samples and Inception V3 model. While the traditional edge detector supposedly identifies all of the sharp changes and does not focus any specific attention to the actual objects, the saliency methods are still focusing on the actual objects for predictions.

that GGIG method retained learned concepts more accurately during post-hoc explanations.

We think that for well-trained deep models, saliency methods no-longer function like an edge-detectors.

**Model Randomization Test:** We also performed a *Model Randomization* test \[109\] (Figure 6.10) to verify the sensitivity of the methods to the model parameters. To this end, we randomly reinitialized the weights and generated post-hoc explanations using the randomized model. The proposed method GGIG is as sensitive as GRAD and IG, suggesting that our method is highly sensitive to model parameters.

6.5.5 Evaluation of Attribution Quality

**Visual Analysis:** We show few sample explanations thresholded at 10% in Figure 6.11. As
Figure 6.10: "Model Parameter Sensitivity" test experiments on ImageNet samples and Inception V3. We show how different methods generate saliency maps when layers of the underlying model (Inception V3) are gradually randomized from top to bottom.

As expected, the edge detector does not pay attention to discriminative regions. Rather, it captures only the sharp changes throughout the image. While GRAD and IG attribute lots of redundant or unexpected parts of the image, GGIG directly points to the most discriminative pixels of the image. Moreover, GGIG mostly avoids sharp changes in the image (see the explanation for the "leopard"), while GRAD and IG are highly susceptible to the edges.

**Quantitative Evaluation:** It is challenging to evaluate an interpretability method because of the lack of ground truth saliency or consensus metrics for proper evaluation. Several studies proposed different measures, such as Remove And Retrain (ROAR) [86], Retain And Retrain (RAR) [18], Accuracy Information Curves, Softmax Information Curves [58] to assess the quality of explanations.

The ROAR approach modifies the dataset by removing the features that received top attribution values from each sample. In practice, if the training data have sufficient (redundant) discriminative features [114], even after removing a significant number of features, the
performance of the retrained model does not drop noticeably. In that scenario, ROAR fails to evaluate the feature relevance correctly. Retain and Retrain (RAR) method resolves the problem by retaining only the critical features instead of removing them. However, both ROAR and RAR methods are time-consuming as they require full retraining of the model.

Dabkowski et al. [115] proposed a metric, called Smallest Sufficient Regions (SSR), based on the notion of the smallest sufficient region capable of correct prediction. However, this metric is not suitable if the model is susceptible to the scale and aspect ratio of the object. Moreover, as this metric depends on rectangular cropping and reports results as a function

---

**Figure 6.11:** Comparison of saliency-focused images thresholded at 10%. GGIG, compared to other methods, is more convincing from the human standpoint because it directly points to the discriminative parts of the image while ignoring redundant and unexpected regions.
of the cropped area, this approach highly penalizes if the saliency map is coherently sparse [58]. Because, in that case, it may span a larger area of the image than the map, which is locally dense, even with the same number of pixels. SSR also imposes a severe challenge because masking creates a sharp boundary between the masked and salient region, causing an out-of-distribution problem for the model.

Kapishnikov et al. [58] proposed two metrics, called Accuracy Information Curve and Softmaxe Information Curve, collectively called Performance Information Curve (PIC), to evaluate the saliency maps with the minimal out-of-distribution setting. We used SIC and AIC to evaluate our method and adopted a similar setup as used in [58]. However, instead of using compressed image size as a proxy information level, we directly computed the entropy of the gray version of the saliency-focused image (test image).

We show the quantitative evaluation results in Figure 6.12. We started with a complete defocused image and gradually added salient pixels to form a saliency-focused image. We measure the entropy of the saliency-focused image relative to the complete blurred image and call it a normalized entropy, which forms the x-axis. We feed saliency-focused images to the original model and report median softmax scores and accuracy calculated over all images, which forms the y-axis. We show the evaluation procedure in Figure 6.2.

While the empirical results are promising, GGIG has some limitations and may need to undergo further reliability tests. First, because of its dependence on gradient ascent optimization GGIG is computationally more demanding relative to the baseline methods. However, all the gradient ascent optimizations from different interpolations can be performed
Figure 6.12: **Left:** Evaluation of saliency methods using Softmax Information Curve (SIC) metric for Inception V3 and ResNet101 models using ImageNet test and validation images respectively. **Right:** Evaluation using Accuracy Information Curve (AIC) metric for the same models and datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Inception v3</th>
<th>ResNet 101</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SIC</td>
<td>AIC</td>
</tr>
<tr>
<td>Gradients</td>
<td>0.337</td>
<td>0.364</td>
</tr>
<tr>
<td>Integrated Gradients</td>
<td>0.409</td>
<td>0.415</td>
</tr>
<tr>
<td>GGIG</td>
<td><strong>0.576</strong></td>
<td><strong>0.523</strong></td>
</tr>
<tr>
<td>Edge Detector</td>
<td>0.384</td>
<td>0.394</td>
</tr>
<tr>
<td>Random</td>
<td>0.201</td>
<td>0.241</td>
</tr>
</tbody>
</table>

in parallel using vector computation. This advantage of parallelization computationally reduces GGIG to a single optimization problem. Second, while empirical performance of GGIG is highly convincing, it is currently lacking deep theoretical insights. Future work shall focus on theoretical foundations of the approach. Further testing is possible in extended studies, such as testing the susceptibility to pointing out unknown spurious signal [595].
6.6 Conclusion

In this chapter, we propose an interpretability method called GGIG that leverages the model’s dynamic behavior in parameter space to enhance the quality of the feature attributions. GGIG starts with a linearly interpolated path as IG and uses gradient ascent to explore how the model interacts with input in the neighborhood. We believe that by looking into the parameter space from different interpolation points, GGIG captures important information about model vs. input interaction. We show the effectiveness of the proposed method through visual inspection and several quantitative metrics across some popular architectures and datasets. As demonstrated, we expect that this work provides insights toward building a more robust interpretability method through model and input interaction.
CHAPTER 7
CONCLUSION AND FUTURE WORK

This chapter provides a brief summary of contributions we made in this dissertation. We also listed potential research directions for future work on interpretability in neuroimaging and other medical domains.

7.1 Summary

In Chapter 1, we have presented an introduction to the context and the research problem we addressed in this dissertation. In Chapter 2, we have reviewed the existing interpretability literature and discussed the current methods, metrics, and practices of “Explainable AI" in general. In Chapter 3, we reviewed over 300 neuroimaging papers that focused on the usage of interpretable machine learning/deep learning. In Chapter 4, we have discussed our first deep learning model designed for efficient transfer learning in neuroimaging domain. Indeed, we showed that how deep learning model can be pretrained using contrastive learning approach and demonstrated its efficacy in synthetic and neurogimaging studies. In Chapter 5, we discussed our proposed deep learning framework that offers end-to-end interpretable pipeline ranging over pretraining, training, model interpretation and explanation validation. In Chapter 6, we discussed our proposed post hoc interpretability method, called Geometrically Guided Integrated Gradients (GGIG), that offers improved deep interpretability compared to integrated gradients—a widely admired method in the literature.

In Chapter 1, we briefly introduced the context of the problem we addressed in this dissertation, mainly the significance and challenges of deep learning and interpretability in
neuroimaging. We described the aims and the objectives and discussed our contributions toward these objectives.

In Chapter 2, we reviewed the significance of interpretability in AI, the philosophy of generating scientific interpretations, and the widely practiced interpretability methods and metrics in the literature.

In Chapter 3, we extensively reviewed and analyzed the current interpretability practices in neuroimaging studies built upon deep learning techniques. We identified some challenges that need to be addressed to make deep learning models valuable in successful human-machine collaborative neuroimaging studies.

In Chapter 4, we developed a pre-trainable deep learning model for neuroimaging. We propose a contrastive self-supervised pretraining method that can pretrain the model via establishing a spatio-temporal objective to maximize the mutual information between the current latent state and future spatial state and between consecutive spatial states. We successfully demonstrated that the proposed model, when pretrained using our pretraining method, outperforms its “w/o pretraining” counterpart both in synthetic and real downstream experiments. For the synthetic experiments, we generated VAR and SVAR datasets. We pretrained our model only based on the VAR dataset and then assessed the performance on the downstream task of VAR vs. SVAR classification. For the real experiments, we used the FBIRN dataset for schizophrenia patients vs. controls classification. The same model was pretrained using HCP subjects. For both cases, we see that the downstream models are directly getting the benefits of pretraining when trained with very few samples.
In Chapter 5, we have proposed a deep learning model, whole MILC, that can be pretrained using the contrastive approach to maximize the mutual information between local and global embedding of the signals. Indeed, we show that DL is able to learn the underlying mechanisms of the disorders better than ML models while maintaining its discriminative power. These representations, as demonstrated, improved downstream results in three different downstream studies with diverse age groups and disorders. We have also shown that our model is interpretable, and some interpretability methods in the literature are able to find the spatiotemporal signature in the underlying signal, which is highly discriminative and informative of the underlying brain dynamics. We have verified the predictability of the model-identified salient regions using our proposed “Retain and Retrain (RAR)” validation framework. We also developed a metric to measure if the explanation is temporally local or distributed. For this metric, we leveraged Wasserstein p-distance with p=1, which we can call the "Earth Mover’s Distance (EMD)." However, we computed EMD from the uniform distribution to characterize the temporal span of the generated explanations.

In Chapter 6, we discussed our post hoc interpretability method, GGIG, that leverages the geometric behavior of the functional space as learned by a model. We also proposed two novel sanity tests, called "\(\sigma - \) perturbation" and "target object sensitivity," that we think post hoc interpretability methods should satisfy. We verified the qualitative and quantitative performance of GGIG on MNIST and ImageNet datasets and compared them against integrated gradients (IG)—a popular interpretability method in the literature. As GGIG is able to identify the contours of the discriminative regions in the input space, we
think GIG will be very effective for different medical imaging tasks where localization as an explanation is a priority.

### 7.2 Future Directions

In current state-of-the-art findings, DL has been proven to be performing well for mental disorder diagnosis and prognosis tasks. However, DL models are black boxes, and their learning mechanism is still not fully understood. Moreover, we must go beyond diagnosis and understand what the model has learned from the data. Importantly, there is a massive chance of clinical misuse with potential adverse outcomes if the model is prematurely designed, trained, tested, and deployed. Recent mental disorder studies paid enormous attention to a growing subfield in AI, traditionally called "Explainable AI (XAI)," to uncover the knowledge the ML/DL models have learned from the brain data. This field has achieved initial success over multiple disorders, which is aligned with the existing literature. In their current forms, however, the findings are still insufficient for individualized treatments in everyday clinical practices. However, the existing best models can play essential roles as supportive tools for individualized diagnosis and prognosis, providing additional validation of the decisions made by human experts. As the field of interpretability in neuroimaging is flourishing rapidly, we believe AI-led solutions will soon be efficient enough to encompass the entire trajectory of diagnosis, prognosis, and treatments. We identified several future directions that we think are worth for further exploration.

**Investigating the true effectiveness of transfer learning in medical imaging:** While
we used self-supervised pretraining to learn from an unlabeled dataset, researchers from medical fields have widely used transfer learning from natural imaging toward different neuroimaging [35] or other medical domains. Raghu et al. [36] showed that transfer learning from natural to medical images did help little with performance. Instead, as the authors surmised, the slight improvement may come from the over-parameterization of the standard models trained on natural images. In line with Chapters 4 and 5, one of the interesting future directions can be investigating the quality and type of knowledge being transferred from the natural image domain to the medical image domain or from one disorder to another disorder.

**Use RAR interpretability framework for disorder progression study:** Our RAR framework, as proposed in Chapter 5, is adaptable and can be helpful for model debugging or data interpretation tasks. For example, the RAR framework can be utilized for a detailed understanding of the stages of progression of Alzheimer’s disease leading to early diagnosis of the disease. When early diagnosis and underlying neuroscientific explanations are feasible, we can probably prevent or decelerate the disease progression by altering the disease progression mechanism.

**Generalized deep interpretability framework:** How can we convert the proposed quantitative validation framework into a domain-supported validation framework? When achieved, we can use XAI to advance our scientific understanding of brains by converting data into knowledge. In line with the RAR framework as described in Chapter 5, one useful area of investigation can be developing a validation framework that is both quantitatively and
scientifically valid for the brain imaging domain [156].

**Using GGIG for interpretability in medical imaging tasks:** As described in Chapter 6, we found GGIG useful for locating the contours of the discriminative regions of the input. We can leverage this interesting property of GGIG for the required interpretable localization in different segmentation, classification, and regression tasks in medical domains.

**Multi-modal fusion and incorporating domain knowledge into interpretability:** We can use multi-modal fusion and incorporate expert knowledge into the learning paradigm. Model debugging is necessary to ensure that the model does not rely on spurious correlations. One can convert model understanding into new knowledge by confirming that the model has learned some aspects of true dynamics and not relying on spurious correlations. One may also use a single modality for developing models but can debug the model using expert features. It is also possible to use the structure-function fusion model for model diagnosis using interpretability.

**Unification of explanation validation:** Earlier studies used an ad-hoc approach for the validation of explanations. The ongoing evaluation approach to model-generated explanations is very subjective. People frequently use only qualitative evaluation with a plausible narrative, but that may not be the case in practice. Some studies used different statistical significance tests to demonstrate the underlying association between marked regions and the predictions. Investigating a unified explanation framework can be an interesting area to work on in the future.
Understanding the underlying causal mechanism: Counterfactual explanations [81] have been proposed in the literature as a means of algorithmic recourse. Investigating counterfactual explanations and the underlying biological/functional mechanism of the disorders can be an interesting future work.

Finding disorder subtypes or data clustering: In almost all earlier investigations, people used model interpretability in classification tasks. Interpretability to find the subgroups within observations is still an unexplored area and can be explored further in the future.
Appendices
Figure 1: RAR employs an SVM to validate FNCs computed using different percentages of the salient input data (5% - 30%) as determined through post hoc explanations of the whole MILC model’s predictions. Salient features are highly predictive compared to the same amount of random features.
REFERENCES


[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep


[58] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai:


[61] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.


[81] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving*


[88] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasub-


on computer vision and pattern recognition, pages 6720–6731, 2019.


[138] Yuan Xiao, Zhihan Yan, Youjin Zhao, Bo Tao, Huaiqiang Sun, Fei Li, Li Yao, Wenjing Zhang, Shah Chandan, Jieke Liu, et al. Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural


[157] Anees Abrol, Manish Bhattacharai, Alex Fedorov, Yuhui Du, Sergey Plis, Vince Calhoun,


[179] Sukrit Gupta, Yi Hao Chan, Jagath C Rajapakse, Alzheimer’s Disease Neuroimaging


Angela Lombardi, Domenico Diacono, Nicola Amoroso, Alfonso Monaco, João Manuel RS Tavares, Roberto Bellotti, and Sabina Tangaro. Explainable deep learning


[202] Huzheng Yang, Xiaoxiao Li, Yifan Wu, Siyi Li, Su Lu, James S Duncan, James C Gee, and Shi Gu. Interpretable multimodality embedding of cerebral cortex using attention


[208] Shangran Qiu, Prajakta S Joshi, Matthew I Miller, Chonghua Xue, Xiao Zhou, Cody


and quantitatively interpretable alzheimer’s disease prediction. 2022.


[237] Tim Hahn, Jan Ernsting, Nils R Winter, Vincent Holstein, Ramona Leenings, Marie


[248] James Zou, David Park, Aubrey Johnson, Xinyang Feng, Michelle Pardo, Jeanelle France, Zeljko Tomljanovic, Adam M Brickman, Devangere P Devanand, José A Luchsinger,


[253] David A Wood, Sina Kafiabadi, Aisha Al Busaidi, Emily Guilhem, Antanas Montvila,


[272] Zülfikar Aslan and Mehmet Akin. A deep learning approach in automated detection of


Chun-Chih Huang, Intan Low, Chia-Hsiang Kao, Chuan-Yu Yu, Tung-Ping Su, Jen-Chuen Hsieh, Yong-Sheng Chen, and Li-Fen Chen. Meg-based classification and grad-cam visualization for major depressive and bipolar disorders with semi-cnn. In


[290] Jiaguang Li, Ying Wei, Chuyuan Wang, Qian Hu, Yue Liu, and Long Xu. 3-d cnn-based


[304] Jingjing Gao, Mingren Chen, Die Xiao, Yue Li, Shunli Zhu, Yanling Li, Xin Dai,


[322] Patrick H Luckett, Luigi Maccotta, John J Lee, Ki Yun Park, Nico UF Dosenbach,


[328] YL Wang, ZJ Zhao, SY Hu, and FL Chang. Clcu-net: Cross-level connected u-


[339] Matthew J Leming, Simon Baron-Cohen, and John Suckling. Single-participant structural similarity matrices lead to greater accuracy in classification of participants than


[357] Sreevalsan S Menon and K Krishnamurthy. Multimodal ensemble deep learning to


[368] Maxim Kan, Ruslan Aliev, Anna Rudenko, Nikita Drobyshev, Nikita Petrashen, Ekaterina Kondrateva, Maxim Sharaev, Alexander Bernstein, and Evgeny Burnaev. Inter-


[381] Obioma Pelka, Christoph M Friedrich, Felix Nensa, Christoph Mönninghoff, Louise


[386] Xinyang Feng, Zachary C Lipton, Jie Yang, Scott A Small, Frank A Provenzano,


[398] Xinyang Feng, Zachary C Lipton, Jie Yang, Scott A Small, and Frank A Provenzano.


[403] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H


[412] Tanu Wadhera and Mufti Mahmud. Computing hierarchical complexity of the brain from electroencephalogram signals: A graph convolutional network-based approach. In 2022 International Joint Conference on Neural Networks (IJCNN), pages 1–6. IEEE,
2022.


[425] Laura de la Fuente, Federico Zamberlan, Hernán Bocaccio, Morten Kringelbach, Gustavo Deco, Yonatan Sanz Perl, and Enzo Tagliazucchi. Temporal irreversibility of neural


[436] Federica Cruciani, Lorenza Brusini, Mauro Zucchelli, Gustavo Retuci Pinheiro,


[470]


[487] Michael Allwright, Hamish Mundell, Greg Sutherland, Paul Austin, and Boris Guen-


[515] Xiaoxiao Li, Yuan Zhou, Nicha C Dvornek, Yufeng Gu, Pamela Ventola, and James S Duncan. Efficient shapley explanation for features importance estimation under uncer-


[547] Alvaro Ulloa, Sergey Plis, and Vince Calhoun. Improving classification rate of schizophren-
nia using a multimodal multi-layer perceptron model with structural and functional

[548] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario


[550] David B Keator, Theo GM van Erp, Jessica A Turner, Gary H Glover, Bryon A Mueller,
Thomas T Liu, James T Voyvodic, Jerod Rasmussen, Vince D Calhoun, Hyo Jong
Lee, et al. The function biomedical informatics research network data repository.

2011.

[552] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson,
Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster,
Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the Human

[553] Yuhui Du, Zening Fu, Jing Sui, Shuang Gao, Ying Xing, Dongdong Lin, Mustafa Salman,
Anees Abrol, Md Abdur Rahaman, Jiayu Chen, et al. NeuroMark: An automated and
adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders.


[566] Natasha Sinegina. Human silhouettes. *Creative Commons https://creativecommons.org/licenses/by/4.0/.*


conference on computer vision and pattern recognition, pages 2818–2826, 2016.