

**DATA DRIVEN APPROACHES TO ADDRESS INACCURATE
NOSOLOGY IN MENTAL HEALTH FROM NEUROIMAGING
DATA**

A Dissertation
Presented to
The Academic Faculty

by

Hooman Rokham

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2023

COPYRIGHT © 2023 BY HOOMAN ROKHAM

**DATA DRIVEN APPROACHES TO ADDRESS INACCURATE
NOSOLOGY IN MENTAL HEALTH FROM NEUROIMAGING
DATA**

Approved by:

Dr. Vince D. Calhoun, Advisor
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Omer T. Inan
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. David Anderson
School of Electrical and Computer
Engineering
Georgia Institute of Technology

Dr. Sergey Plis
Department of Computer Science
Georgia State University

Dr. Shella Keilholz
College of Engineering and Emory School
of Medicine/ Biomedical Engineering
Georgia Institute of Technology

Date Approved: December 11, 2023

Dedicated to my beloved Haleh and our precious son, Leo, the lights of my life.

ACKNOWLEDGEMENTS

My sincere gratitude goes to Dr. Calhoun, an exceptional supervisor whose mentorship has been a steadfast beacon during this transformative journey. His boundless patience, kindness, and profound wisdom have enriched my understanding beyond measure. The guidance he provided played a pivotal role in paving the way for the successful completion of this research. I want to convey my heartfelt thanks to Dr. Plis, whose unwavering support has been a pillar of strength throughout the challenges encountered in this research. I also extend my appreciation to the esteemed members of my thesis committee, Dr. Anderson, Dr. Keilholz, and Dr. Inan for their profound insights that greatly contributed to shaping this work. I would like to express a special recognition for Haleh, my dear friend, lab companion, and beloved wife. Her steadfast presence has served as a guiding light and a source of strength throughout this journey. Her unwavering faith in me has been a continuous wellspring of inspiration, and her limitless kindness and unwavering support have been constant companions through every high and low of this extraordinary adventure.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF SYMBOLS AND ABBREVIATIONS	Error! Bookmark not defined.
SUMMARY	xvi
CHAPTER 1. Introduction	1
1.1 Neuroimaging in Brain and Mental Disorders: Unlocking the Mysteries	3
1.1.1 Structural MRI Analysis: Decoding the Brain's Blueprint	3
1.1.2 Functional MRI Analysis: Unveiling the Brain's Dynamic Patterns	4
1.1.3 Multimodal Imaging Analysis: Unraveling the Complexity of Mental Disorders	6
1.2 Psychosis Categorization Heterogeneity: Decoding the Enigma of Mental Illness	8
1.3 Label Noise in Classification: Challenges in Mental Health Classification	9
CHAPTER 2. A Data-driven Approach for Stratifying Psychotic and Mood Disorders Subjects Using Structural Magnitude Resonance Imaging DataSetting up a new chapter	13
2.1 Introduction	14
2.2 Materials and Methods	15
2.2.1 B-SNIP Dataset	15
2.2.2 Image Processing	17
2.2.3 Method Description	18
2.3 Results	20
2.4 Discussion	27
2.5 Conclusion	27
CHAPTER 3. Addressing Inaccurate Nosology in Mental Health: A Multilabel Data Cleansing Approach for Detecting Label Noise From Structural Magnetic Resonance Imaging Data in Mood and Psychosis Disorders	29
3.1 Introduction	29
3.2 Materials and Methods	32
3.2.1 Bipolar–Schizophrenia Network on Intermediate Phenotypes Dataset	32
3.2.2 Preprocessing	39
3.2.3 Method Description	39
3.3 Results	47
3.4 Discussion	62
3.5 Conclusion	66

CHAPTER 4. Evaluation of boundaries between mood and psychosis disorder using dynamic functional network connectivity (dFNC) via deep learning classification Setting up a new chapter	67
4.1 Introduction	68
4.2 Materials and Methods	71
4.2.1 Bipolar–Schizophrenia Network on Intermediate Phenotypes Dataset	72
4.2.2 Preprocessing and ICNs extraction	75
4.2.3 dFNCs and Exemplar clustering	76
4.2.4 Group differences and State Occupancies	77
4.2.5 Classification	77
4.2.6 Deep model Classification	78
4.2.7 SVM Classification	80
4.2.8 Classification setup	80
4.2.9 Classification analysis	81
4.3 Results	82
4.3.1 Intrinsic connectivity networks and dynamic functional connectivity	82
4.3.2 Classification results	86
4.3.3 Classification analysis	88
4.4 Discussion	91
4.5 Conclusion	96
CHAPTER 5. A Deep Learning Approach for Psychosis Spectrum Label Noise Detection from Multimodal Neuroimaging Data	97
5.1 Introduction	98
5.2 Materials and Methods	99
5.2.1 Data Information, Preprocessing, Intrinsic Connectivity Network Extraction	99
5.2.2 Convolutional Neural Networks (CNNs)	100
5.2.3 Classification Tasks and Noise Estimation	102
5.2.4 Groups Comparison	102
5.3 Results	103
5.4 Discussion	106
CHAPTER 6. Neuroimaging Data Informed Mood and Psychosis Diagnosis: Label Noise-Robust Ensemble Deep Multimodal Framework for Neuroimaging Data	109
6.1 Introduction	110
6.2 Materials and Methods	113
6.2.1 Data information	113
6.2.2 Preprocessing and ICNs extraction	114
6.2.3 Classification tasks	116
6.2.4 Deep learning base models	117
6.2.5 Ensemble method	118
6.2.6 Label noise analysis	120
6.3 Results	122
6.4 Discussion	133
6.5 Conclusion	136

CHAPTER 7. Conclusion	138
REFERENCES	141

LIST OF TABLES

Table 2-1 – Demographic of B-SNIP dataset	17
Table 2-2 – ANOVA test statistics on Clusters, DSM, Biotype Categories	24
Table 3-1 – B-SNIP Dataset Demographic	34
Table 3-2 – Number of samples for Biotype and DSM-IV groups	35
Table 3-3 – Medicine status of the subjects in B-SNIP dataset	37
Table 3-4 – B-SNIP dataset different parameter settings for the scanners	38
Table 3-5 – Shared label noise using consensus voting	50
Table 3-6 – Area of containing voxels obtained by statistical significant test on the relabeled data ● and original data ◇	60
Table 4-1 – B-SNIP Dataset Demographic	74
Table 4-3 – Classification Tasks	78
Table 4-4 – States’ Occupancy	84
Table 4-5 – Number of significant pair obtained from ANOVA statistical test	85
Table 4-6 – Deep model classification results	87
Table 4-7 – Total number of significantly different FC pairs of misclassified subjects with their original and predicted group	89
Table 4-8 – Number of significant FC pairs of misclassified subjects with their original and predicted group	90
Table 5-1 – Classification report	105
Table 6-1 – Demographic Table	114
Table 6-2 – Classification Tasks	117
Table 6-3 – Classification Results	123
Table 6-4 – Classification Performance Using Different Datasets.	127
Table 6-5 – Label Noise Group Difference Analysis	131

LIST OF FIGURES

Figure 2-1	– Diagram of the proposed method	18
Figure 2-2	– Contingency matrices of clustering and available categories (DSM-IV, biotype, Schizo-Bipolar Scale) of the B-SNIP dataset. Each row shows a type, and each cell shows the percentage and number of subjects of its type. A) Shows the contingency matrix of DSM-IV and clusters. B) Shows the contingency matrix of the biotype and the clusters. B) Shows the contingency matrix of the Schizo-Bipolar Scale and the clusters. The estimated distribution of the Schizo-Bipolar Scale score for each cluster shown in Figure 2-3 cluster 3 grouped more subjects with the lowest Schizo-Bipolar Scale score, and cluster 1 contained the largest number of subjects with the highest Schizo-Bipolar Scale score. Other clusters included subjects with a different score range, mostly higher score ones.	23
Figure 2-3	– Schizo-Bipolar Scale score distribution for each cluster	25
Figure 2-4	– Visualization of extracted features in 2-dimensional space using tSNE.	25
Figure 2-5	– Gray matter map results of voxel-wise ANOVA statistical test. Each cluster only showed regions if they were significantly different from other clusters, and their mean values on those regions were higher than other clusters.	26
Figure 3-1	– Visual summary and workflow of different aspects of the proposed method. Preprocessing was done using SPM on structural MRI data collected at multiple sites. Hyper-parameters optimization and parameter selection among choosing the best dimension reduction were done using grid search. Then univariate feature selection using ANOVA was selected as best approach and used for the dimension reduction approach, applied for dimension reduction on preprocessed data and subsampling was done on dataset to handle imbalanced classes. Cross validated classification/voting filtering was applied to resampled data. Then noisy labels were identified based on their votes, cleansed dataset (exclude noise subjects) and unlabeled dataset (noisy subjects with dropped labels) feed to supervised to evaluate and obtain the new suggested labels. Combining cleansed data and data with new suggested labels make cleansed and relabeled data. For generalization, using a convergence approach, subsampling, cross-validated classification voting filtering and relabeling	41

performed iteratively on cleansed and new labeled data till we identified a specific acceptable amount of label noise in the dataset.

- Figure 3-2 – a) Shows random samples from handwriting digits dataset. The numbers of 8,1,3,5 were selected because of their similarities and common features they have; and b) The mean over all samples for each digit. 45
- Figure 3-3 – Deep model convolutional neural network ResNet architecture. 3D structural MRI scans are seen by the convolution layer followed by batch normalization and a rectifier unit layer. The output of these layers goes through max pooling layer and then go through Residual block of containing a series of 3D convolutional, batch normalization and non-linear rectifier (ReLU) layers 47
- Figure 3-4 – Shows effect of different proportion of label noise on accuracy of SVM classifier and improving accuracy using data cleansing approach. Corrupting data by shuffling labels of instances and adding synthetic label noises to ground truth decreases the performance of the predictive model. By applying data cleansing classification filtering method, we could identify noisy labels and improve the model performance. Even when 80% of labels in dataset are noisy, model accuracy was boosted from 0.38 in the noisy dataset to 0.9 in the cleansed dataset. The effect of label noise on splitting dataset also shows evaluating data cleansins is challenging when the ground truth are not available and may mislead during interpretation. 48
- Figure 3-5 – Average accuracy of classification filtering cross validation part after 5 iterations. In each iteration, label noise is identified, and new labels are assigned based on similarities between them and cleansed instances. The performance of the model increases after each iteration. 49
- Figure 3-6 – Top: Confusion matrices show label noise subjects related to Biotype or DSM-IV categories separately after convergence to 90% of cleansed data using consensus voting. Middle: Confusion matrices show shared label noise subjects among Biotype and DSM-IV categories after convergence. About 50% of Biotype proband 2 (B2) and 45% of Biotype proband 3 (B3) were relabeled as normal control. Among Biotype proband 1 (B1), 81% of instances identified as noisy label were relabeled as Biotype proband 2 (B2) and Biotype proband 3 (B3). 81% of normal control noisy subjects were also relabeled as Biotype proband 2 (B2) and Biotype proband 3 (B3). In DSM-IV category, 71% of bipolar proband noisy subjects were relabeled as healthy controls, 51

whereas 64% and 32 % of healthy controls were relabeled as bipolar proband and schizoaffective proband respectively. Among schizophrenia label noise subjects, 38% were relabeled as healthy controls and 34% as schizoaffective proband. Bottom: Confusion matrices show the result of predicting labels of relatives using updated labels after the above analysis. Most of the relatives are labeled as Biotype3 (B3) and normal controls in the Biotype categories or as healthy controls in the DSM-IV category.

- Figure 3-7 – Left panel shows t-SNE 2D-projection of original dataset with label noise and right panel shows 2D-projection using new suggested labels. Top: 2D-projection using Biotype labels. Bottom: 2D-projection using DSM-IV labels. In both categories of Biotype and DSM-IV, affinities between data points shows subjects in different group of disease and healthy control overlap in noisy data. After identifying label noise and relabeling, the similarity between data points is more obvious and subjects with the same labels are close together. Embedded data into 2D space using original labels does not support the fact that subjects labeled based on their similarities. It is hard to interpret how groups differentiate from each other using original label, because there is considerable overlap between subjects. Cleansed dataset with new suggested label shows there is gradient pattern in both DSM-IV and Biotype categories from healthy/normal control to most severe cases of Biotype B1 or schizophrenia probands contain other mild cases in between. 54
- Figure 3-8 – Gray matter map results of voxel wise t-tests between 4 Biotype groups after data cleansing using classification voting filtering. Top row) NC vs. B1, NC vs. B2, NC vs. B3. Bottom row) B1 vs. B2, B1 vs. B2, B2 vs. B3. Gray matter contrast between normal controls and Biotype probands shows they have differences with different levels in overlapped regions after cleaning and relabeling individuals. Gray matter differences between normal control and Biotype proband 1 (B1) have strongest separation among the other comparison tests. The gray matter contrast is lower between healthy controls and Biotype proband 3 (B3). Biotype proband 2 (B2) and Biotype proband 3 (B3) show the fewest differences. 56
- Figure 3-9 – Gray matter map results of voxel wise t-tests between 4 Biotype groups on given labels. Top row: NC vs. B1, NC vs. B2, NC vs. B3; Bottom row: B1 vs. B2, B1 vs. B2, B2 vs. B3. Gray matter contrast between normal controls and Biotype probands only shows significant difference in some regions between normal controls and Biotype proband 1 (B1) using original labels. Gray 57

matter density group differences between other groups do not show differences.

- Figure 3-10 – Gray matter map results of voxel wise t-tests between 4 DSM-IV groups after data cleansing using classification voting filtering. Top row) BPD vs. HC, HC vs. SADP, HC vs. SZP. Bottom row) BPD vs. SADP, BPD vs. SZP, SADP vs. SZP. Gray matter contrast between healthy controls and DSM-IV probands shows they have differences with different levels in some overlapped regions after cleaning and relabeling individuals. Gray matter difference between healthy control and bipolar proband (BPD) has strongest separation among the other. The gray matter contrast shows differences between healthy control and schizophrenia and schizoaffective probands as well. Also, group differences could be observed between the bipolar proband group vs. schizophrenia and bipolar proband group vs. schizoaffective proband groups. However, schizophrenia and schizoaffective groups did not separate after data cleansing and relabeling. 58
- Figure 3-11 – Gray matter map results of voxel wise t-tests between 4 DSM-IV groups on given labels. Top row: BPD vs. HC, HC vs. SADP, HC vs. SZP. Bottom row) BPD vs. SADP, BPD vs. SZP, SADP vs. SZP. Gray matter contrast was only significant in some areas between healthy controls and schizophrenia proband using the original labels. Group differences between other groups were not found. 59
- Figure 3-12 – ROC plot of 12 binary classification tasks using Biotype and DSM-IV labels on cleansed dataset after 1st iteration of our data cleansing approach before convergence. After one iteration of classification filtering on data and training deep model using cleansed dataset the accuracy improved about 20% from 0.65 to 0.79 on an average of 6 binary classification task using DSM-IV labels and by 23% from 0.60 to 0.74 on 6 binary classification task using Biotype labels. 61
- Figure 3-13 – Accuracy boxplot of 12 binary classification tasks on deep model on cleansed dataset after 1st iteration. Red are boxplot accuracy of stratified cross validation using original label. Blue boxplot shows the accuracy of different stratified cross validation run on cleansed dataset after a single classification voting filtering. The figure shows accuracy improved on cleansed dataset after a single iteration of classification filtering. 62
- Figure 4-1 – Method diagram of our proposed method. After preprocessing, steps, group independent component analysis using Neuromark pipeline was conducted on resting-state fMRI data and 53 intrinsic 72

connectivity networks (ICNs) extracted, and spatial maps and time courses were estimated for each subject using the back-reconstruction method. Using the window approach, dynamic FNC features were estimated and k-means clustering was applied to dFNC features to estimate FNC states. Statistical analysis and group differences were conducted on dFNC features across diagnoses of DSM biotype categories. From cluster medians of each group and regression analysis, β coefficients were estimated as the features for classification. Comprehensive and various binary and multiclass classification tasks were conducted using the support vector machine (SVM) and the proposed deep models. Statistical analyses were performed between misclassified and correctly classified subjects to extract potential hidden dFNC biomarkers which might add more homogeneity between groups.

Figure 4-2	– Number of subjects for each group of DSM-IV and Biotype categories of B-SNIP dataset. The percentage in front of each group shows how each proband group diagnosed as other categories.	75
Figure 4-3	– Deep classification model architecture.	80
Figure 4-4	– Dynamic states of 4 groups of DSM-IV and Biotype categories averaged across each state’s windowed dNFC. The result of the ANOVA statistical test and identified significant pair of connections are highlighted in the last row.	83
Figure 4-5	– Classification accuracy of the total 22 tasks, including using deep and SVM models. For each task, the average accuracies obtained from cross-validation runs of SVM and deep models are shown. The tasks sorted from the highest accuracy (left side) belong to the DSM-IV binary task between the normal control group and schizoaffective group to the lowest accuracy (right side) of the biotype 4-way classification task.	87
Figure 4-6	– Confusion matrices of 3-way and 4-way classification tasks of DSM-IV and biotype categories obtained from deep model. The confusion matrices were normalized over the true (given) labels and by rows.	88
Figure 4-7	– The connectogram (chord diagram) of group difference between misclassified subjects and their original and suggested group by the model. Rows are the original (given labels and the columns are the suggested labels. Each cell belongs to a pair of given label group and predicted label group based on the relation of their differences with the misclassified subjects. Red links shows significant different FC pair between misclassified subjects and	89

their original groups and green links significant different FC pair between misclassified subjects and their suggested correctly classified groups.

Figure 4-8	– 3D tetrahedron scatter plot of the probability values of correctly classified (left) and misclassified subjects (right) for DSM group(top) and Biotype group (bottom). The dots close to the vertices have a higher probability for that class from the model.	91
Figure 5-1	– Architecture of our proposed multimodal deep models. Top, 3D convolutional neural network for sMRI data, and bottom, 1D convolutional neural network for fMRI data.	101
Figure 5-2	– Probability distribution obtain by the model.	103
Figure 5-3	– Average of static functional connectivity for different groups. The static FNCs average from clean and noisy subjects of the normal control (NC) group are on the left panel, and the average of cleansed and noisy samples of the Schizophrenia group are on the right panel.	104
Figure 5-4	– Cell-wise static functional connectivity group difference of different groups. Upper triangle is cell-wise subtraction between functional connectivity of each group and lower triangle shows FDR corrected significant cells obtained from Welch’s t-test at significant level of 0.05	106
Figure 6-1	– Method Diagram and Convolutional Neural Network Models. The diagram illustrates the methodology employed in this study, incorporating subsampling and a repeated cross-validation approach to enhance the evaluation of each sample using multiple models trained on different subsets of the data. By generating random subsets of the data, we ensure diverse training samples for each model. The predictions from these models are then combined using a voting approach, enabling the determination of final labels for the subjects in both the unimodal and multimodal approaches. This methodology enhances the robustness and generalizability of the classification process by leveraging the collective knowledge of multiple models.	120
Figure 6-2	– The figure presents the confusion matrices resulting from the application of the ensemble deep multimodal approach to classify subjects across 11 different tasks. Each matrix represents the performance of the classification model by displaying the counts or percentages of true positive, true negative, false positive, and false negative predictions for each diagnostic group. The matrices provide a comprehensive visual representation of the model's	125

accuracy and misclassification patterns, offering valuable insights into the categorization outcomes based on neuroimaging features.

- Figure 6-3 – Performance comparison of the models with and without the inclusion of additional data. The results demonstrate improved classification accuracy when more control samples are added using the FBIRN dataset. The model trained solely on the B-SNIP datasets shows comparable performance to the model utilizing all datasets, indicating minimal impact of dataset variation on the model's effectiveness 127
- Figure 6-4 – Group differences in static functional connectivity (FC) pairs for the 4-way classification task. Comparison between all data and clean (label noise removed) data shows an increase in the number of significant FC pairs, indicating improved discrimination between groups. 132
- Figure 6-5 – Group differences in static functional connectivity (FC) pairs for the 4-way classification task. Comparison between noisy data and the corresponding clean group, as well as noisy data and the suggested clean group, reveals distinctive patterns of FC differences, highlighting the presence of label noise and the potential for improving classification accuracy. 132

SUMMARY

Label noise poses a significant challenge in machine learning, where erroneous labels are assigned to instances. These inaccuracies stem from diverse sources such as inadequate information, human error, and subjective interpretation, among others. In the realm of supervised learning, particularly in classification methods, where models are trained using labeled data to predict labels for unseen samples, label noise emerges as a critical concern. It not only adversely impacts classification accuracy but also complicates model intricacies.

Despite the strides in computational power and deep learning techniques, the issue of learning from noisy datasets with erroneous labels remains notably formidable, especially in sensitive domains like medical applications. This challenge becomes even more daunting in medical imaging, where mislabeled data can lead to incorrect diagnoses and potentially harmful outcomes.

In the specialized field of psychiatry, this problem reaches a pinnacle of complexity. Existing categorizations of psychosis are marred by limitations such as unreliability, heterogeneity, and the absence of a definitive ground truth. These factors exacerbate the challenges in addressing label noise in psychiatric data.

In light of these challenges, this research aims to achieve two pivotal objectives. The first goal is to develop robust frameworks and algorithms for detecting incorrect labels within datasets. These methodologies will strive to enhance data homogeneity by employing semi-supervised auto-labeling approaches. By identifying and rectifying

mislabeled instances, the research aims to mitigate the adverse effects of label noise, particularly in medical imaging and psychiatric datasets.

The second goal delves into the realm of biomarker discovery related to mood and psychosis disorders. The research endeavors to construct a comprehensive framework capable of unveiling latent structural and functional patterns within neuroimaging data. These patterns, once identified, could serve as vital biomarkers, offering profound insights into the understanding and treatment of mood and psychosis disorders. By providing alternative diagnostic and classification methods, this research seeks to revolutionize the approach to these complex mental health conditions, thereby advancing the field significantly.

CHAPTER 1. INTRODUCTION

The human brain, with its staggering complexity comprising over 100 trillion interconnected synapses and neurons, serves as the nexus of our thoughts, emotions, and behaviors, making it one of the most intricate networks in existence. Unraveling the enigma of the brain's structure, functions, and their intricate ties to clinical disorders stands as a monumental challenge in the realm of scientific inquiry.

Recent years have witnessed an intensified focus on this challenge through multidisciplinary research, where the convergence of various scientific fields has led to translational studies. These studies have a singular goal: to identify elusive biological markers from the vast landscape of neuroimaging data. The aim is profound - to diagnose, detect vulnerabilities, and gauge the severity of mental disorders with unprecedented precision, ultimately paving the way for more effective treatments. This pursuit has captivated the attention of researchers and scholars worldwide, underscoring the urgency and significance of the endeavor.

Adding to this momentum are the remarkable advancements in machine learning techniques. These cutting-edge tools have found application in the intricate terrain of psychiatric disorders, showing promising outcomes. The integration of machine learning methodologies has propelled the field forward, offering valuable insights and novel perspectives [1]. Multiple studies on structural and functional brain imaging provided various findings related to different brain disorders [2].

In this research, our focus is keenly directed towards the analysis of structural and functional imaging modalities, specifically delving into the realms of gray matter structural MRI and resting-state functional MRI. These technologies serve as our lenses into the intricate workings of the human brain, allowing us to explore the subtle nuances that underlie mood and mental disorders. Furthermore, we confront the vexing challenge posed by the current categorization of psychoses. In the world of artificial intelligence, where precision and accuracy are paramount, the existing categorization methodologies face profound challenges, leading to a quest for innovative solutions.

As this chapter unfolds, we embark on a comprehensive journey. We traverse related works, delving into the vast body of knowledge that precedes our endeavors. We confront the complexities of current psychosis categorization, dissecting the challenges that impede progress. Efforts to surmount these challenges come under scrutiny, revealing the innovative strategies that researchers employ to transcend existing limitations.

Lastly, we delve into the intricate realm of label noise phenomena. In the context of machine learning classification, this phenomenon emerges as a critical obstacle. The presence of label noise, akin to a disruptive static in an otherwise clear signal, poses significant challenges. We explore the nuances of this challenge and its potential interconnection with the intricacies of current psychosis nosology, shedding light on the symbiotic relationship between machine learning and the understanding of mental disorders.

In essence, this chapter serves as a gateway to a world where cutting-edge technology converges with the complexities of the human mind. Through rigorous exploration and

innovative thinking, we strive to unravel the mysteries that lie at the intersection of neuroscience, artificial intelligence, and psychiatric disorders, in the relentless pursuit of advancements that could transform the landscape of mental health care.

1.1 Neuroimaging in Brain and Mental Disorders: Unlocking the Mysteries

1.1.1 Structural MRI Analysis: Decoding the Brain's Blueprint

The realm of structural MRI (Magnetic Resonance Imaging) has been a fervent battleground of exploration, particularly in the quest to understand various brain and mental disorders [2]. One of its most triumphant conquests has been the identification of structural biomarkers for Alzheimer's disease, a progressive neurodegenerative condition [2]–[4]. Groundbreaking studies employing support vector machines (SVM) and manual feature extraction have not only enabled diagnosis and prognosis but also opened avenues for therapeutic advancements in Alzheimer's cases [5], [6]. Recent strides in deep learning techniques have further refined this diagnostic process, especially in multi-class classification tasks, underscoring the potential of artificial intelligence in unraveling the complexities of neurodegenerative diseases [7], [8]. The main reason for this success is that Alzheimer's disease is progressive neurodegenerative.

However, the terrain becomes notably rugged when delving into affective and non-affective psychoses such as Schizophrenia, bipolar disorder, and schizo-affective disorders. Unlike Alzheimer's, these conditions lack a well-established neuropathological foundation, making the task of identifying biomarkers considerably more challenging. Binary classification attempts, particularly in distinguishing individuals with bipolar disorder from control groups, have yielded accuracies ranging from 59.50% to 73.00%, showcasing the

intricate nature of these disorders and the gaps in our understanding of their pathophysiology [1], [9]–[12]. The findings suggest that the pathophysiology of this disorder is still not well understood, and it is difficult to diagnose bipolar disorder.

Intriguingly, meta-analytical approaches have revealed notable differences in gray matter volume between Schizophrenia patients and control subjects [13]. Unmedicated Schizophrenia patients exhibit severe deficits in hippocampal volume, pointing toward significant neural variations in these conditions [2], [13]. Yet, the application of machine learning techniques for binary classification tasks in Schizophrenia has yielded a wide spectrum of accuracy, spanning from 56% to an astonishing 96.3% [14]–[17]. In an attempt to disentangle the complexities, studies have ventured into differentiating Schizophrenia from bipolar disorder using SVM classification on structural MRI features. Despite reported average classification accuracies of 76%, 66%, and 61% between control vs. Schizophrenia, bipolar vs. Schizophrenia, and control vs. bipolar respectively, inconsistencies persist due to limited sample sizes, demanding further validation [12], [18].

In this intricate landscape of neural imaging, these findings underscore both the remarkable progress and the challenges that persist in our endeavor to decode the mysteries of the human brain, underscoring the need for robust methodologies and continued research to bridge these gaps in our understanding.

1.1.2 Functional MRI Analysis: Unveiling the Brain's Dynamic Patterns

In recent years, resting-state functional MRI (fMRI) features have emerged as a pivotal tool in diagnosing psychoses, marking a paradigm shift in our understanding of

brain disorders. Unlike its structural counterpart, functional MRI delves into the dynamic, real-time neural activities, illuminating intricate patterns of the brain's functioning.

Both resting-state and task-based functional MRI studies have showcased a remarkable leap in accuracy when compared to structural classification methods, especially in binary diagnosis tasks such as Schizophrenia vs. control and bipolar disorder vs. control groups [1]. The versatility of resting-state functional MRI becomes evident in Schizophrenia diagnosis, with reported accuracies spanning from 62% to an impressive 100%, albeit with variances due to sample sizes and classification settings, emphasizing the nuances involved in data interpretation, and not to be generalizable across other studies.

In the realm of major depression disorder (MDD) diagnosis, the potential of resting-state functional MRI becomes even more apparent. A balanced dataset of 40 subjects served as the canvas for unveiling the complexities of MDD. Utilizing region-based resting-state functional MRI features and employing a Support Vector Machine (SVM) classifier, researchers achieved a staggering accuracy of 95%, showcasing the precision of this method in capturing the subtleties of neural patterns associated with depression [19]. Further expanding the horizon, a three-class classification involving MDD, Schizophrenic, and control groups demonstrated the robustness of feature extraction techniques on resting-state functional MRI data, coupled with SVM classification, achieving an accuracy milestone of 84.2% [20].

These breakthroughs in functional MRI analysis underscore the vital role of dynamic neural activity in understanding psychoses. With advancements continuing to reshape our methodologies, functional MRI not only sheds light on the intricacies of brain disorders

but also offers a beacon of hope in the pursuit of accurate diagnoses and, ultimately, effective treatments.

1.1.3 Multimodal Imaging Analysis: Unraveling the Complexity of Mental Disorders

The exploration of mental disorders through neuroimaging lenses transcends the boundaries of a singular modality, leading us into the intricate realm of multimodal imaging analysis. While individual imaging modalities offer valuable insights, they are inherently limited in capturing the full spectrum of complexities inherent in psychiatric conditions [21]–[23]. Multimodal studies, however, serve as a beacon of hope, illuminating a path towards a more profound understanding of these disorders.

In the pursuit of enhanced diagnostic accuracy, the fusion of diverse data streams holds immense promise. The synergy between different modalities can unravel an array of related biomarkers, shedding light on various disorders and their classifications [1], [24], [25]. Multimodal analysis, a burgeoning field, stands out for its ability to refine diagnosis in challenging cases. Techniques such as data fusion and the creation of sophisticated multimodal models are becoming central to the endeavor of unraveling the intricate tapestry of mental illnesses [22].

Examining Schizophrenia, a condition marked by its clinical heterogeneity, provides a poignant illustration of the power of multimodal approaches. When compared to unimodal studies, combining resting-state functional MRI and structural MRI in a multimodal analysis significantly boosts model performance [26]–[29]. Accuracy, once languishing at 52% with structural MRI and 78% with functional MRI in unimodal

classifications, soared to an impressive 87% in the multimodal setting, underscoring the profound impact of integrating information streams [26], [27].

Delving even deeper, cutting-edge research ventures into the realm of multimodal deep learning. A groundbreaking study fused structural, functional, and genomic data to differentiate Schizophrenia from control cases. The results were nothing short of revolutionary, showcasing a significant improvement in accuracy and highlighting the potential of advanced machine learning techniques in addressing the complexities of mental disorders [30].

However, challenges persist. In the landscape of psychosis categorization, overlapping symptoms and subgroup diagnoses pose formidable obstacles. Single-case studies, while valuable, often overlook these complexities and the inherent heterogeneity across groups, demanding a nuanced approach [1]. Additionally, the current literature on other psychosis and mental disorders remains limited. Robust validations are imperative before the multimodal biomarker can be seamlessly integrated into clinical practices, ensuring both accuracy and reliability in diagnostics.

In the evolution of machine learning techniques, traditional models like Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) held sway in the past. However, with the surge in computational power, deep learning techniques have assumed a pivotal role in neuroimaging data analysis [31], [32]. These advancements herald a new era, where the fusion of multimodal data and the synergy of sophisticated algorithms promise not just progress but a profound transformation in our comprehension of mental disorders.

1.2 Psychosis Categorization Heterogeneity: Decoding the Enigma of Mental Illness

Within the labyrinthine realm of psychiatry, the quest for a biological foundation of mental illnesses has been a formidable challenge. Existing categorization methods, epitomized by the APA Diagnostic and Statistical Manual (DSM), grapple with inherent limitations, rendering them less than entirely valid [33], [34]. Diagnoses, particularly for conditions as complex as Schizophrenia and bipolar disorder, predominantly rely on symptom-based assessments, an approach plagued by unreliability.

Compounding this challenge is the intricate web of overlapping symptoms and biological markers prevalent across different DSM diagnoses. These shared elements extend beyond clinical manifestations, infiltrating biological realms encompassing disease risk genes, structural and functional brain metrics, electrophysiological patterns, and cognitive functional deficits. The clinical and neurobiological heterogeneity inherent in these disorders further confounds the diagnostic landscape.

To bridge this chasm between symptoms and underlying biological realities, a multifaceted approach emerges. Integrating clinical data with genomics, alongside other pertinent patient information, such as brain biomarkers, emerges as a beacon of hope. This holistic perspective not only refines the identification of valid disease subtypes but also elevates the precision of subject categorization, ultimately enhancing treatment outcomes [35]. This amalgamation of diverse data streams provides a richer, more nuanced understanding of mental disorders, transcending the limitations of traditional psychiatric diagnoses.

Yet, the conundrum persists. The dilemma lies in reconciling these biological insights with the inherent shortcomings of established psychiatric diagnostic frameworks. Applying biological classifications within the confines of traditionally flawed diagnostic methods raises profound questions about the productivity of such endeavors. Unraveling these inaccuracies stands as a crucial endeavor, driven by a profound interest in the overarching mission: to aid those afflicted by mental illness [36].

Stepping back, it becomes evident that the rigidity of categorical diagnostic approaches, focused on predicting specific diseases, blinds us to the subtleties inherent in mental health. The emphasis on continuous measures emerges as an enlightened perspective. Beyond categorical classifications, this nuanced approach recognizes the multifaceted nature of mental disorders. Continuous measures not only augment our diagnostic capabilities but also offer a deeper, more nuanced understanding of the complex interplay between symptoms and underlying biological realities [37].

In this profound exploration, the field of psychiatry finds itself at a crossroads, where the fusion of biological insights and innovative diagnostic paradigms holds the potential to illuminate the path forward. Through this lens, the enigma of mental illness begins to unravel, revealing opportunities for enhanced diagnosis, precise categorization, and ultimately, transformative treatment modalities.

1.3 Label Noise in Classification: Challenges in Mental Health Classification

In the expansive landscape of artificial intelligence, classification stands tall as a pivotal tool, guiding the process of predicting the class of new samples in the realm of machine learning. The bedrock of this predictive prowess lies in the foundation of

supervised classifier algorithms, which meticulously learn from data, meticulously analyzing the provided labels to make informed predictions for unseen samples. However, the veracity of these predictions hinges profoundly upon the reliability of the dataset itself.

Herein lies the challenge: the whisper of label noise, an ever-present murmur that disrupts the harmony between data features and their assigned classes. Label noise, in essence, shrouds the true relationship between the characteristics of an instance and its rightful classification, creating a haze that muddles the path to accurate predictions [38]–[41]. In an ideal scenario, a pristine, label noise-free dataset would be the linchpin of accurate predictions. However, in the realm of mental health, such purity is a rare commodity. The intricacies of current diagnostic strategies, marred by inaccuracies and questionable validity, cast a shadow upon the cleanliness of the data, allowing the intrusion of label noise into the very core of our analyses [38], [42].

Understanding the impact of label noise on classification accuracy has been a subject of profound inquiry. Previous studies have illuminated this critical connection, revealing the delicate balance upon which accurate predictions teeter [38], [43]. Amidst this uncertainty, neural networks have emerged as stalwart defenders, showcasing their resilience in the face of label noise. Yet, the tale takes a twist in the realm of deep learning. Here, the thirst for vast datasets is unquenchable, for deep learning models, while robust, demand a bountiful reservoir of clean training data to function optimally [44], [45].

In this intricate dance between data and prediction, label noise emerges not merely as an obstacle, but as a revealing reflection of the complexities inherent in mental health data. As we navigate these murmurs, understanding their nuances becomes paramount.

Each disturbance, each whisper of uncertainty, propels the field forward, urging researchers to delve deeper, work smarter, and innovate ceaselessly. In this pursuit, the elucidation of the relationship between label noise and classification accuracy serves not just as a challenge but as a beacon, guiding the way toward more precise, more reliable, and ultimately, more transformative insights in the realm of mental health diagnostics.

The challenges posed by the existing limitations in psychosis categorization, including concerns about its reliability, heterogeneity, and the absence of a valid ground truth, represent a significant hurdle in the field of psychiatry. Addressing these challenges is pivotal for enhancing categorization and nosology in psychiatry.

This research aims to tackle these issues through two primary objectives. Firstly, our goal is to develop robust frameworks and algorithms for handling label noise, enabling the identification and correction of incorrect labels within the dataset. Through an innovative semi-supervised labeling approach, we aim to maximize homogeneity across groups by implementing auto-labeling techniques. Secondly, we strive to create a comprehensive framework capable of unveiling hidden structural and functional patterns within neuroimaging data. By identifying specific biomarkers associated with mood and psychosis disorders, our objective is to enhance understanding and potentially revolutionize treatment approaches. This research endeavors to explore alternative methods of diagnosis and classification that could prove invaluable in the context of mental health.

To achieve these objectives, our investigation delves into both structural and resting-state functional MRI data across mood and psychosis disorders. We employ a range of

approaches, including supervised, semi-supervised, and unsupervised methodologies, to analyze these modalities comprehensively. Our study involves a meticulous comparison between the existing psychosis nosology and potential data-driven approaches, juxtaposed against the current gold standard categorization and continuous measures within different modalities.

In the subsequent sections, we will provide a detailed account of the methodologies and investigations undertaken to address these pressing challenges.

**CHAPTER 2. A DATA-DRIVEN APPROACH FOR
STRATIFYING PSYCHOTIC AND MOOD DISORDERS
SUBJECTS USING STRUCTURAL MAGNITUDE RESONANCE
IMAGING DATASETTING UP A NEW CHAPTER**

Psychotic disorders such as schizophrenia and bipolar disorder are difficult to classify because they share overlapping symptoms. Deriving biomarkers of illness using structural MRI dataset are essential because they may lead to improved diagnosis. Previous studies typically predict the diagnosis labels using supervised classifiers that rely on truly labeled dataset. Mislabeled subjects may increase the complexity of the predictive model and may impact its performance. In this work, we address the problem of inaccurate diagnosis labeling of psychotic disorders using a data-driven approach. We performed dimension reduction using PCA on the vectorized images and then k-mean clustering on the components. An ANOVA statistical significance test performed after clustering based on DSM, biotype, and clustering of images. We evaluate our method on a structural MRI dataset, which over 900 subjects labeled using DSM-IV and biotypes. Subjects were grouped into 5 clusters using our method, and each cluster includes all types of patients. However, we found statistically significant group differences in brain regions across 5 clusters, while for DSM and biotype, there were not statistically different. Also, our result shows the performance of the predictive model improved significantly using data-driven labels. Our method shows underlying biological changes associated with mental illness may be identified by studying and considering features of the brain imaging data, and

annotating brain imaging data using a data-driven approach may eventually lead to improved diagnosis and advance drug discovery and help patients.

2.1 Introduction

Identifying a biological basis for mental illness for psychotic disorders such as schizophrenia and bipolar disorder are challenging and. Finding distinctive patterns for mood disorders to classify is difficult because they share overlapping symptoms overlap in their clinical signs and biological measures like risk genes, structural and functional brain measures, electrophysiology, and functional cognitive deficits [33], [34]. A third diagnostic entity, schizoaffective proband, was created to capture individuals who have both psychosis and mood symptoms, but these are still known to be imperfect as there is no biological basis for the illness. This difficulty motivates the study of other aspects of the disorder such as the history of mental illness in the family. The use of brain imaging e.g. structural MRI (sMRI) data promises to provide biological information that may be helpful in identifying illness.

To derive biomarkers of illness, previous studies typically predict the diagnosis labels using structural MRI data. Machine learning approaches often involve supervised algorithms that rely on a dataset with accurate labels; however, their performance may be affected due to inaccurate labeling, as mentioned previously. It is unlikely to have a completely pure and label noise-free dataset in almost every real-world application [38]. The field of mental health is not exceptional, especially given the fact that current diagnostic approaches are inaccurate, and the validity of the current stratification methods is questionable. Computer-aided diagnosis (CAD) tools can help with this task and may

lead to discovering new biomarkers for better classification. Lots of studies used different techniques e.g., deep neural network, gaussian graphical model (GGM) to classify and identify differences between mood disorders [12], [46]–[48]. Other studies in computer-aided diagnosis (CAD) field for brain imaging used unsupervised segmentation for the diagnosis of Alzheimer's disease [49].

In our experiment, we performed classification task using a large MRI data set including over 900 individuals. Classification results using a state-of-the-art convolutional neural network ResNet model provided relatively low accuracy. To address the mislabeling issue, we utilize an approach to first partition the data into more homogeneous clusters, followed by testing between groups and extracting features as potential biomarkers. Our proposed method used an unsupervised approach to obtain shared features in structural MRI and having within-subjects homogeneity of each category by suggesting new labels and subcategories.

2.2 Materials and Methods

2.2.1 B-SNIP Dataset

We used the bipolar-schizophrenia network on intermediate phenotypes (B-SNIP-1) [50], [51] structural imaging (structural MRI) dataset. The B-SNIP is a multiple-site consortium that was formed to study various intermediate phenotypes across psychotic disorders and to find their relation to the phenomenologically characteristic derived from DSM-IV [51]. Psychotic probands studied in B-SNIP categorized by DSM-IV into schizophrenia, schizo-affective disorder, and bipolar disorder with psychosis. Alongside with DSM-IV, the subjects categorized by observed neurobiological heterogeneity defined

groups, called biotypes (biotype1, biotype2, and biotype3) [52]. Biotype group B1 manifest cases with impaired cognitive control and poor sensorimotor function, group B2 was characterized by impaired cognitive control but exaggerated sensorimotor response, and group B3 presented near normal cognitive and sensorimotor functions [53]. The subjects also were rated by experts using another scaled called Schizo-Bipolar Scale that is a 10 point scale that rates based on overall psychotic illness, overall extent of manic syndrome as a proportion of the overall illness duration, and type of mood symptoms [50]. Scores are in a range between 0 and 9 that 0 are most bipolar-like and 9 is most schizophrenia-like disorder. Healthy controls in this dataset did not have PANSS and Schizo-Bipolar Scale scores. The dataset that we used in this study included 901 subjects aged 15-65 years. More detail about data reported in the demographic Table 1. There are different 3T scanners (GE Signa, Philips Achieva, Siemens Allegra, and Siemens Trio) across different sites for acquiring three-dimensional structural MRI data.

Table 2-1 – Demographic of B-SNIP dataset

	DSM-IV								Biotype							
	BPD		SADP		SZP		HC		B1		B2		B3		NC	
	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>
Size	173	19.2	131	14.5	238	26.4	359	39.8	145	16.1	181	20.1	216	23.9	359	39.8
	<i>Me</i>	<i>Std</i>	<i>Me</i>	<i>Std</i>	<i>Me</i>	<i>Std</i>	<i>Me</i>	<i>Std</i>	<i>Me</i>	<i>Std</i>	<i>Me</i>	<i>Std</i>	<i>Me</i>	<i>Std</i>	<i>Me</i>	<i>Std</i>
Age	36.2	13.15	35.62	11.91	34.44	12.34	37.24	12.44	35.43	13.17	35.2	11.97	35.12	12.53	37.24	12.44
Schizo-Bipolar Scale	1.07	1.06	5.11	1.51	7.94	1.14	-	-	6.2	2.7	5.15	3.24	4.27	3.24	-	-
PANSS																
Positive	12.95	4.47	18.09	5.21	16.64	5.61	-	-	16.52	5.83	15.88	5.73	15.33	5.18	-	-
Negative	12.1	3.75	15.54	4.67	16.74	6.04	-	-	16.33	5.69	15.28	5.35	13.79	5.14	-	-
General	28.64	7.96	34.51	8.91	32.16	8.75	-	-	31.58	8.95	32.44	9.01	30.95	8.55	-	-
Total	53.71	13.56	68.09	16.25	95.5	17.13	-	-	64.42	17.56	63.56	17.01	60.60	16.21	-	-

2.2.2 Image Processing

Collected structural MRI data (1mm isotropic MPRAGE) of subjects [51], [54] were preprocessed via SPM (<https://www.fil.ion.ucl.ac.uk/spm/>) using a unified approach [55]. Preprocessing steps included tissue classification, bias correction, image registration, spatial normalization, and resliced to 2x2x2mm. Also, to identify and avoid the impact of the segmentation outlier, the unsmoothed gray matter density (GMD) images were then correlated with gray matter template. Outliers have been corrected if possible or removed otherwise. Moreover, to identify problematic scans, we calculated correlation among all gray matter segmentation maps and removed those that had lower correlation (<0.7). We

also scaled structural MRI data and covaried for each site to reduce the impact of different sites with different scanners. The 3D gray matter images have 121×145×121 size.

2.2.3 Method Description

In this work, we propose the use of a data-driven approach to the impact of inaccurate labels on our predictive model. We applied our approach to voxelwise gray matter segmented images from a structural MRI dataset. We first performed dimension reduction using PCA on the vectorized images. Then, we were using cross-validated k-mean clustering on the components by dividing data into two sets of training and prediction. We fit the k-mean model the training set, and after obtaining the model, we used the model to predict the label (cluster) for each subject. To avoid bias, we repeated the cross-validation step multiple times. Then, we obtain the centroid of the clusters for different iteration and find the euclidian space to find the best match of the clusters for each repetition and made the labels of the subjects identic. Then, we used to have different numbers of labels obtained by our approach and assigned used majority voting to get final labels for each subject. We implemented our method in python 3.7 utilizing scikit-learn library [56]. The best optimal number of clusters identified using the elbow method. An ANOVA statistical significance test performed after clustering based on DSM, biotype, Schizo-Bipolar Scale, and clusters of images. Figure 2-1 shows the flow chart diagram of our proposed method.

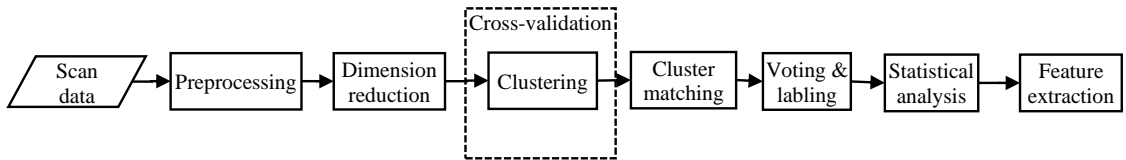


Figure 2-1 – Diagram of the proposed method

2.2.3.1 Dimension reduction

Course-of-dimensionality phenomena occur in the neuroimaging field because the number of features (voxels) is significantly larger than the number of observations (subjects). Dimensionality reduction approaches are used to remove redundant variables and only keep features that are informative in showing differences of samples to mitigate this issue. In this study, we used Principal component analysis (PCA) [57], which constructs relevant features by transforming raw voxels of structural MRI data into a subset of features called principal components, which are uncorrelated features [58].

2.2.3.2 Clustering

To categorize subjects using extracted features from dimension reduction, we used clustering groups similar subjects into the same cluster and put nonsimilar subjects into other clusters. In this study, we used k-mean clustering, which is a clustering method that minimizes within-cluster sum-of-squares and separate samples in n different groups (clusters) of equal variance. We used the grid search method using a range of 2 to 8 for the number of clusters. To generalize, we did stratified cross-validated clustering in the way of dividing the dataset into two parts. 1. Training set (fitting) the k-mean model, and 2. The validation set to predict labels based on the k-mean model. We repeated this step multiple times randomly.

2.2.3.3 Cluster Matching

Then to match different k-mean models in various iterations of the cross-validation step, we calculated the Euclidian distance of centroids of different models and considered

the closest centered as similar clusters of different iterations. Then, we update the labels based on the matched clusters, which can be comparable.

2.2.3.4 Voting and Labeling

Using reconstructed labels of different iteration for each subject, we applied majority voting and assigned the label to the subject that repeated more than the others in cross-validation clustering.

2.2.3.5 Statistical analysis

We then used Analysis of variance (ANOVA) to obtain which regions of brain images were statistically different using clustering labeling and were not statistically different using DSM-IV, biotype labeling. We also used voxel wise statistical group differences between clusters to obtain distinctive features of clusters.

2.3 Results

Using the grid search and elbow method in the dimension reduction and clustering steps, we set the number of clusters and the number of components to 5. We performed 2-fold repeated cross-validated clustering and repeated it three times. Using the centroids of k-mean models in these three iterations, we obtained the most similar clusters in different iterations by calculating the Euclidian distance between centroids of features. The minimum values indicated the most similar clusters for different iterations. Then cluster labels changed according to the first iteration cluster centroid as base cluster labels. Consequently, cluster labels assigned to the subjects also changed.

After applying majority voting on subject's cluster labels, subjects grouped into 5 clusters, and various diagnosis and biotype categories overlapped between clusters. The contingency matrices of clustering in Figure 2-2 show how the subjects grouped in different clusters. Each row shows the percentage of a group belonging to different categories of Biotype, DSM-IV, and Schizo-Bipolar Scale. Regarding DSM-IV classes, cluster 2 included most of the healthy control by 25.9% among other groups. About 29.8% of schizophrenia and 22.9% of schizo-affective probands placed in cluster 1. About 25.4% of bipolar disorder grouped in cluster 3. Cluster 1 contained most of the subjects in schizo-affective and schizophrenia. Cluster 2 included most of the subjects of probands that having similar features related to most healthy control. Cluster 3 shared most features between healthy control and bipolar disorder. Cluster 3 contains 19.1% of bipolar probands and 18.1% of schizophrenia probands. Cluster 5 included 23.4% of the healthy control and 20.6% of the schizo-affective proband. In the Biotype perspective, 30.3% of the biotype1 grouped in cluster 1 and about 25.4%, 26.9% and 25.9% of biotype2, biotype3 and healthy control clustered in cluster 2. cluster 3 contained 23.7 of healthy control and 20.8% of biotype3 and 19.3% of biotype2. Cluster 4 included 29.7% of biotype1, and cluster 5 contained 23.4% of normal control. The contingency matrix of Schizo-Bipolar Scale shows the highest proportion of subjects with the Schizo-Bipolar Scale score between 0 and 3 categorized into cluster 3 and the highest proportion of subjects with the highest Schizo-Bipolar Scale score, between 8,9 grouped in cluster 1 and cluster 2.

To show new findings using our method, we applied different analyses on the data after clustering. Each cluster includes all types of patients, and useful information can extract using voxel-based statistical test ANOVA on structural MRI data. To mitigate

multiple comparison test, we applied false discovery rate (FDR) controlling on the calculated p-values. We found statistically significant group differences across 5 clusters in numerous regions including cerebellum, hippocampus, hippocampal, cingulate, entorhinal cortex, nucleus accumbens, superior temporal gyrus, precuneus, paracentral gyrus, frontal pole, transverse temporal gyrus, and postcentral gyrus while for DSM and biotype there were no statistically significant results in those regions. More information about the ANOVA statistical result is in Table 2-2.

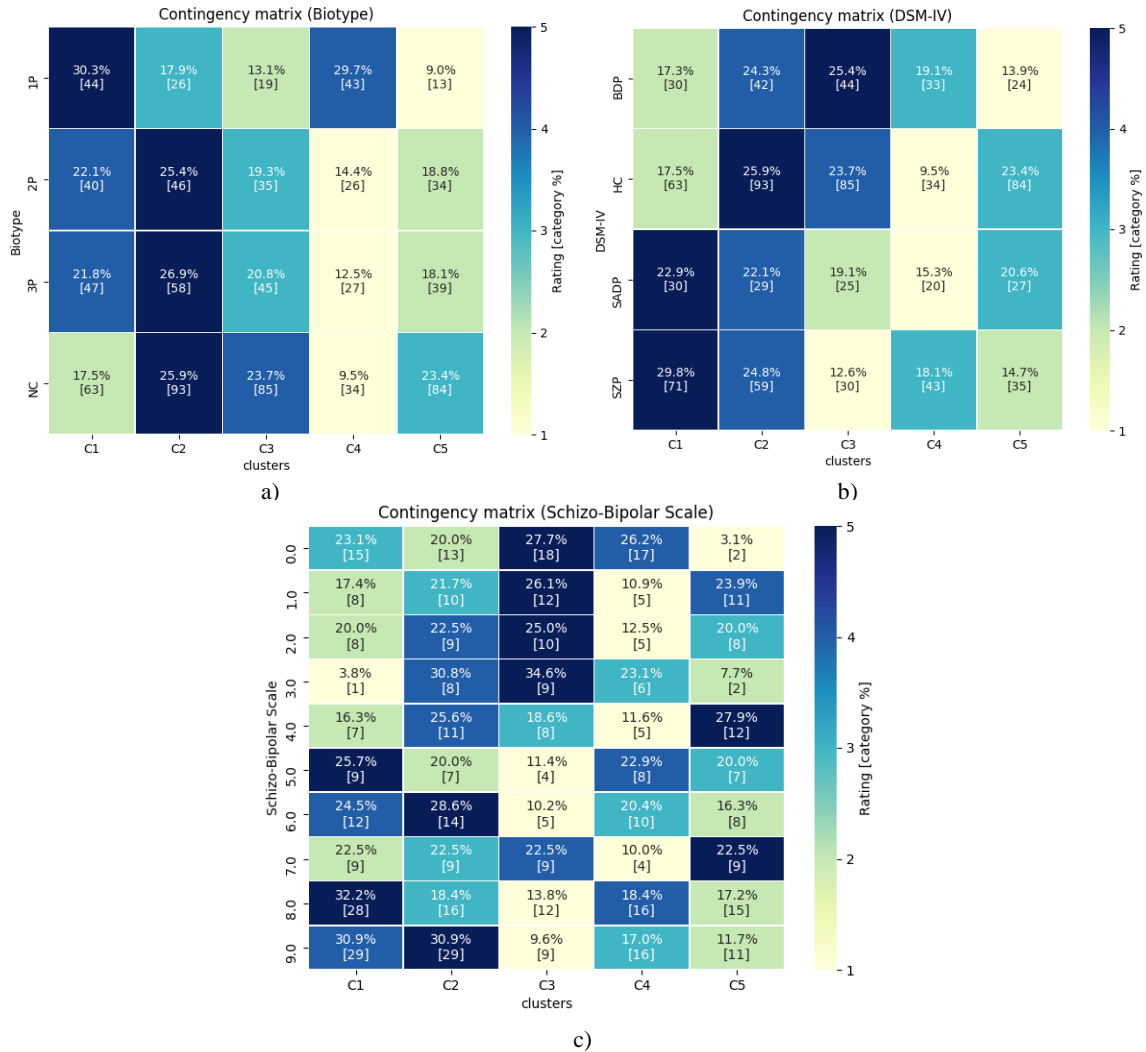


Figure 2-2 – Contingency matrices of clustering and available categories (DSM-IV, biotype, Schizo-Bipolar Scale) of the B-SNIP dataset. Each row shows a type, and each cell shows the percentage and number of subjects of its type. A) Shows the contingency matrix of DSM-IV and clusters. B) Shows the contingency matrix of the biotype and the clusters. B) Shows the contingency matrix of the Schizo-Bipolar Scale and the clusters. The estimated distribution of the Schizo-Bipolar Scale score for each cluster shown in Figure 2-3 cluster 3 grouped more subjects with the lowest Schizo-Bipolar Scale score, and cluster 1 contained the largest number of subjects with the highest Schizo-Bipolar Scale score. Other clusters included subjects with a different score range, mostly higher score ones.

Table 2-2 – ANOVA test statistics on Clusters, DSM, Biotype Categories

		Clustering Statistics		DSM Statistics		Biotype Statistics	
		<i>F</i>	<i>P-Value</i>	<i>F</i>	<i>P-Value</i>	<i>F</i>	<i>P-Value</i>
Left Frontal Pole		2.72	0.02	0.31	0.26	0.06	0.97
Left lateraloccipital		5.91	<0.001	0.98	0.39	2.61	0.05
Postcentral		9.41	<0.001	1.62	0.18	1.22	0.3
Left Temporal	Transverse	4.1	<0.001	2.22	0.08	1.34	0.25
Right Accumbens	Nucleus	19.07	<0.001	0.72	0.53	1.38	0.24
Right Frontal	Medialorbital	7.53	<0.001	2.16	0.09	2.46	0.06
Right Paracentral		18.88	<0.001	1.55	0.19	1.98	0.11
Right Precuneus		16.93	<0.001	1.66	0.17	2.29	0.07
Right Superior Temporal		10.86	<0.001	2.36	0.06	2.33	0.07
Right Superior Temporal		8.69	<0.001	2.31	0.07	2.36	0.06

We also used T-distributed stochastic neighbor embedding (tSNE) projection to show how our proposed clustering method separated subjects in comparison with DSM-IV and biotype categories. Figure 2-4 shows tSNE projection of different types. Univariate feature selection using ANOVA F-value between cluster label and features applied on

preprocessed images to obtain 100 highest score features among the total number of voxels. Extracted features fitted to tSNE to project in data 2-dimensional space. We used the same projection with different coloring to show differences between different categories.

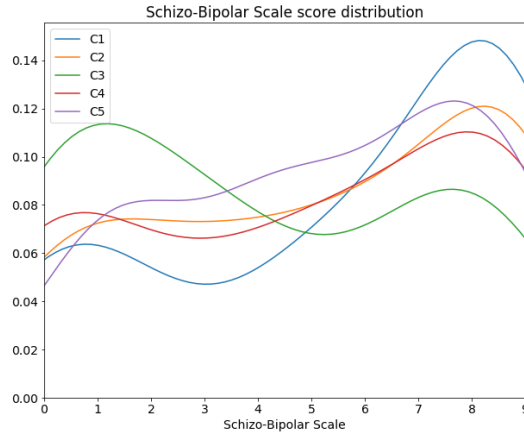


Figure 2-3 – Schizo-Bipolar Scale score distribution for each cluster

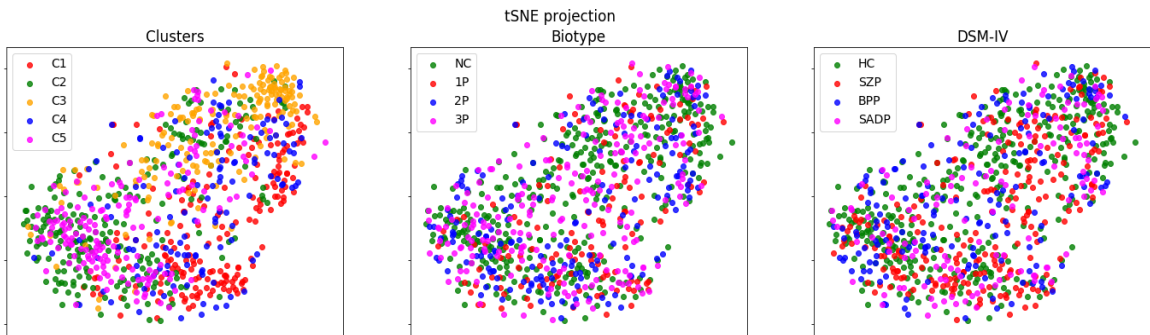


Figure 2-4 – Visualization of extracted features in 2-dimensional space using tSNE.

Voxel-wise ANOVA test applied on non-zero voxels between each category and FDR correction performed for multiple comparisons correction. Among 124,591 voxels, 101,524 of the voxels were significant different using the ANOVA statistical test. However, after applying FDR correction, for DSM-IV, only 2,350 voxels were found significantly different, and for biotype categories, only 3,538 number of voxels were

significant differences with a significant level of 0.05. Considering the condition that the significant difference only occurs in clustering approach and excluding the significant shared voxels of DSM-IV and biotype categories, lead having 97,093 voxels for only the clustering approach. We then calculated the mean of each voxel vector of the subjects of different clusters and obtained images for each cluster in the way that each cluster's images only contain the value of the mean of the voxels if they have the highest mean value for that voxel in comparison with other groups. Otherwise, the values for that voxel set to 0. Figure 2-5 shows the brain regions of belonging clusters.

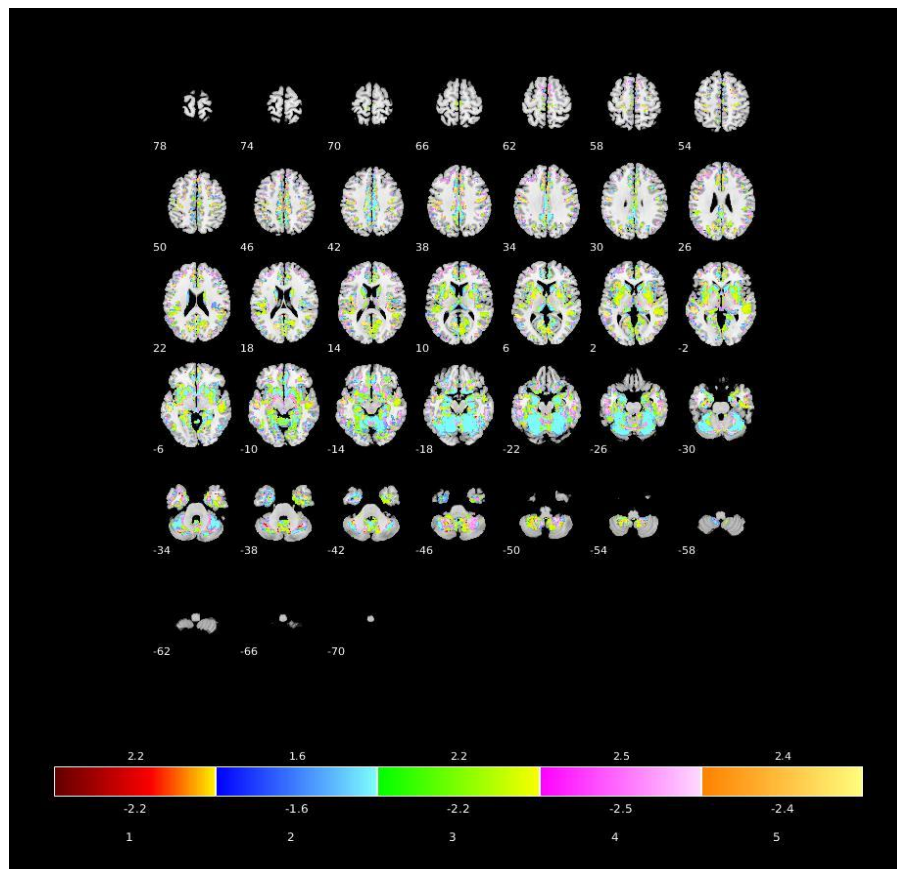


Figure 2-5 – Gray matter map results of voxel-wise ANOVA statistical test. Each cluster only showed regions if they were significantly different from other clusters, and their mean values on those regions were higher than other clusters.

2.4 Discussion

To avoid the effect of dimension reduction on the underlying pattern and to prevent losing some essential part of the underlying trend, in this study, we used grid search with different numbers of components. However, we noticed that using large number of components in comparison with the number of samples will lead to an overfitting problem and should be avoided. Lack of reliable grand truth of brain imaging data in the field of computer-aided diagnosis tool for psychosis and mood disorder is challenging.

In the clustering step, we used k-mean clustering algorithm, which aims to minimize the within-cluster sum-of-squares criterion which is based on assumptions that clusters are convex and isotropic. Using other clustering approaches with different assumptions and benchmarking other strategies of clustering left for future studies.

Although some clustering evaluation metrics were indicating, there are overlaps in clusters, but our method grouped similar subjects based on extracted features from their structural MRI data and made more clear boundaries between subjects of different clusters were not available before clustering and using original DSM-IV or even biotype labeling. But this may raise the question that are the biomarkers extracted using structural MRI data related to mood disorder diagnosis? There are a couple of studies that showed that there are differences between control and some proband groups in their structural MRI data.

2.5 Conclusion

Studying structural MRI and annotating brain imaging data using a data-driven approach and classifying the subjects based on the features of their brain imaging data may

enhance the diagnosis and treatment of patients. The use of clustering to first stratify the data followed by testing within these more homogeneous subgroups enhances sensitivity to the existing diagnostic categories. The proposed approach is a step forward that may provide clues to the underlying biological changes associated with mental illness and may eventually lead to improved diagnosis.

CHAPTER 3. ADDRESSING INACCURATE NOSOLOGY IN MENTAL HEALTH: A MULTILABEL DATA CLEANSING APPROACH FOR DETECTING LABEL NOISE FROM STRUCTURAL MAGNETIC RESONANCE IMAGING DATA IN MOOD AND PSYCHOSIS DISORDERS

In this chapter, we outline our research conducted on T1 structural magnetic resonance imaging data from 1493 individuals comprising healthy control subjects, patients with psychosis, and their unaffected first-degree relatives. Our study focuses on implementing a data cleansing approach to identify label noises, which are subsequently reevaluated and relabeled in a semi-supervised manner. Simulation results showed that our method was highly accurate in identifying label noise. Both diagnostic and biotype categories showed about 65% and 63% of noisy labels, respectively, with the largest amount of relabeling occurring between the healthy control subjects and individuals with bipolar disorder and schizophrenia as well as in unaffected close relatives. The extraction of imaging features highlighted regional brain changes associated with each group. This approach represents an initial step toward developing strategies that need not assume that existing mental health diagnostic categories are always valid but rather allows us to leverage this information while also acknowledging that there are misassignments.

3.1 Introduction

Psychiatry has struggled with identifying a biological basis for mental illness. Current categorization approaches, including the DSM, are not entirely valid [38]–[41].

Diagnosis of mental illness, such as schizophrenia and bipolar disorder, is typically based on unreliable symptom-based measures. It is also known that among different DSM diagnoses, there is a considerable overlap not only in their clinical symptoms but also in biological measures, including disease risk genes, structural and functional brain measures, electrophysiology, and cognitive functional deficits [33], [34], [46]. Additional challenges include the debate over the validity of additional mixed diagnostic categories such as schizoaffective disorder [46]. Unreliable self-report information regarding symptoms makes the process of diagnosis more complicated and leads to incorrect labeling. Understanding brain biomarkers in psychiatry may help to diagnose and treat mental disorders more effectively. Integrating clinical data with genomics and other patient information such as brain biomarkers helps to better define valid disease subtypes and/or categorize subjects more accurately and improve treatment outcomes [35]. The challenge is that applying biological classification while using traditional psychiatric diagnoses that lack inherent validity as ground truth is unlikely to prove productive.

Identifying and prioritizing diagnostic categorization is a major challenge in psychiatry that, if not carried out correctly, translates into inaccuracies in diagnostic labeling of biological data such as medical imaging. Addressing these inaccuracies (which we refer to as label noise in a diagnostic classification problem setup) is an important topic of great interest that serves the ultimate goal of helping patients [36]. The application of artificial intelligence can be leveraged to help with this task and to achieve better results. Multiple studies have shown that there are significant differences among individuals with schizophrenia, individuals with bipolar disorder, and control subjects regarding their

structural magnetic resonance imaging (sMRI) data, especially in the cerebellum and temporal lobe regions [12], [47], [48], [59].

Among different forms of artificial intelligence approaches, classification—the process of predicting the class of new samples—has been broadly used in the realm of machine learning. To compute a prediction for an unseen sample, a supervised classifier algorithm first learns from data based on the provided labels. As such, the reliability of the dataset plays a vital role in the performance of the classification models. Therefore, if there is label noise in the dataset, the prediction accuracy will decrease. However, having a wholly pure and label noise-free dataset is unlikely in mental health given the fact that current diagnostic strategies are inaccurate, and diagnoses are of questionable validity [38], [42]. Noise in this context refers to anything that obscures the relationship between the features of an instance and its class [38], [60]. In addition to label noise, data quality issues (e.g., low signal-to-noise ratio, head motion) may corrupt the relationship between the features of an instance and its class as well. We performed quality control and preprocessing steps to help mitigate this, but we cannot rule it out completely. Previous work has shown how label noise may affect classification accuracy [38], [43]. Studies show that neural networks are robust in handling label noise in data [44]; however, deep learning models require large datasets [45], and the minimum amount of needed clean training data increases with an increase in the label noise level [44].

To improve categorization or nosology in psychiatry, it is essential to address the challenges in the existing categorization. One approach is to consider existing categories as noisy (i.e., containing mislabeled samples) and develop approaches to eliminate, or at least reduce, the impact of this noise during a classification task. In this work, we identify

cases where there is biological evidence that pushes against an existing categorization, that is, cases where gray matter data look similar for subjects who are categorized differently or look different for groups that fall within the same diagnostic category.

Our proposed method involves using a voting approach to estimate which individuals are labeled (diagnosed) incorrectly, which is called class noise or label noise. Then we relabel noisy subjects such that subjects are more similar to each other in the new group. We then repeat these steps until we identify a specific acceptable amount of label noise in the dataset.

By applying our method, we identify shared features of existing categories in sMRI data and regroup subjects who were mislabeled or violated the assumption of homogeneity within groups and suggest the new labels similar to others, which leads to having homogeneity within subjects of each category at the end. Results show that there are homogeneous subsets that fall both within and between existing categories, and we suggest that future studies should focus more on these aspects of the available data.

3.2 Materials and Methods

3.2.1 Bipolar–Schizophrenia Network on Intermediate Phenotypes Dataset

We analyzed the Bipolar–Schizophrenia Network on Intermediate Phenotypes (B-SNIP) sMRI dataset [51], [54]. B-SNIP is a multisite National Institutes of Health–funded consortium of investigators that collected multiple brain imaging and assessment measures for stable patients within three psychotic disorders (schizophrenia, schizoaffective disorder, and bipolar disorder with psychosis). The B-SNIP dataset used in this study

included 912 subjects after quality control assessment. sMRI 3D acquisitions were carried out on 3T scanners: Signa (GE Healthcare, Wauwatosa, WI: University of Chicago, Chicago, IL, and Harvard Medical School, Boston, MA), Achieva (Philips, Best, Netherlands: UT Southwestern Medical Center, Dallas, TX), and Allegra and Trio (Siemens, Erlangen, Germany: Olin Institute of Living, Hartford Hospital, Hartford, CT, and Maryland Psychiatric Research Center, University of Maryland, Baltimore, MD) [53], [61]. High-resolution isotropic T1-weighted magnetization prepared rapid acquisition gradient-echo (MPRAGE) sequences were acquired following the Alzheimer's Disease Neuroimaging Initiative (ADNI) protocol [61], [62]. Subject demographics and the mean and standard deviation of the positive and negative syndrome scale (PANSS) for different categories are reported in Table 3-1 which also includes the mood status of each category. PANSS are not available for control subjects.

Table 3-1 – B-SNIP Dataset Demographic

	DSM-IV								Biotype							
	BPD (N=176)		SADP (N=134)		SZP (N=240)		HC (N=362)		B1 (N=147)		B2 (N=185)		B3 (N=218)		NC (N=362)	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Size	176/	19.3	134/	14.6	240/	26.3	362/	39.6	147/	16.1	185/	20.2	218/	23.9	362/	39.6
	912	0%	912	9%	912	2%	912	9%	912	2%	912	9%	912	0%	912	9%
Gender (Male)	61/1	34.6	56/1	41.7	164/	41.7	162/	44.7	77/1	52.3	90/1	48.6	114/	52.2	162/	44.7
	76	6%	34	9%	240	9%	362	5%	47	8%	85	5%	218	9%	362	5%
Race																
African American	34		52		107		99		79		57		57		99	
American Indian	-		-		1		3		1		-		-		3	
Asian	5		2		5		15		-		7		5		15	
Caucasian	131		73		112		231		57		114		145		231	
Native Hawaiian	-		-		-		1		-		-		-		1	
Multiracial / Mixed Race	2		6		8		6		6		3		7		6	
Other Race	4		1		7		6		4		4		4		6	
Unknown / Missing	-		-		-		1		-		-		-		1	
	<i>Mea</i>	<i>Std</i>	<i>Mea</i>	<i>Std</i>	<i>Mea</i>	<i>Std</i>	<i>Mea</i>	<i>Std</i>	<i>Mea</i>	<i>Std</i>	<i>Mea</i>	<i>Std</i>	<i>Mea</i>	<i>Std</i>	<i>Mea</i>	<i>Std</i>
	<i>n</i>		<i>n</i>		<i>n</i>		<i>n</i>		<i>n</i>		<i>n</i>		<i>n</i>		<i>n</i>	
Age	35.9	13.1	35.8	12.0	34.5	12.3	37.1	12.4	35.5	13.1	35.3	11.9	35.1	12.6	37.1	12.4
		4	8	1	1	3	5	4	2	4	2	1	1	4	5	4
PANSS																
Positive	13.0	4.55	18.1	5.2	16.6	5.69	-	-	16.5	5.83	16.0	5.85	15.2	5.19	-	-
	2		7		9				6		6		9			
Negative	12.1	3.8	15.5	4.63	16.7	6.03	-	-	16.3	5.66	15.2	5.35	13.8	5.17	-	-
	9		2		6				1		8		7			
General	28.7	3.07	34.5	8.82	32.2	8.71	-	-	31.6	8.88	32.5	8.94	31.0	8.58	-	-
	8		4						3		2		1			
Total	54.0	13.6	68.1	16.0	65.6	17.1	-	-	64.4	17.4	63.8	17.0	60.1	16.2	-	-
		8	9	9	2	6			8	7	2	2	3	1		

Data were categorized and labeled using two different approaches. This includes both standard clinical diagnosis based on the DSM-IV and also neurobiological heterogeneity defined groups, called Biotypes [52]. Each Biotype category contained individuals with all DSM psychosis categories and vice versa (see Table 3-2). According to [52], “Biotypes”

(biologically distinctive phenotypes) refer to neurobiologically distinct subgroups of psychosis cases independent of clinical phenomenology that differentiated people with psychosis from healthy controls. Biotype group B1 manifest cases with impaired cognitive control and poor sensorimotor function, group B2 was characterized by impaired cognitive control but exaggerated sensorimotor response, and group B3 presented near normal cognitive and sensorimotor functions [53]. DSM categories included bipolar disorder (BPD) with psychosis, schizoaffective disorder (SADP), schizophrenia (SZP) and also healthy control (HC) subjects. In this study, we separately evaluated the DSM and Biotype labeled data.

Table 3-2 – Number of samples for Biotype and DSM-IV groups

DSM-IV	Biotype	# of individuals
BPD	B1	25
	B2	57
	B3	94
SADP	B1	33
	B2	48
	B3	53
SZP	B1	89
	B2	80
	B3	71
HC	NC	362

The B-SNIP dataset also includes 581 patients’ relatives (Biotype patient relatives) which consist of 193 bipolar relatives (BPDR), 152 schizoaffective disorder relatives (SADPR) and 236 schizophrenia relatives (SZPR). Within the Biotype categorization, 147

of them are relatives of Biotype group 1 (BR1), 191 subjects are relatives of Biotype group 2 (BR2), 116 subjects are relatives of Biotype group 3 (BR2). Also, among 581 relative subjects only 82 of them have PANSS available. 26 bipolar proband relatives with average PANSS 52.3 +/- 14.5, 28 schizoaffective proband relatives with average PANSS 58.14 +/- 17 and 28 schizophrenia proband relatives with average PANSS 58.28 +/- 17.26.

The medication status of the individuals reported in this dataset is shown in Table 3-3. About 95% of bipolar disorder subjects, 95% of the schizoaffective proband and 92% of schizophrenia patients reported that they were using psychotropic medications.

Table 3-3 – Medicine status of the subjects in B-SNIP dataset

Medicine type	Reported	Total numbers	BPD		SADP		SZP		HC	
			<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Any Psychotropic medications	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
	NO	370	13/176	7.39%	7/134	5.22%	13/240	5.42%	337/362	93.09%
	YES	524	161/176	91.48%	127/134	94.78%	223/240	92.92%	13/362	3.59%
Antipsychotic	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
	NO	430	44/176	25.00%	16/134	11.94%	20/240	8.33%	350/362	96.69%
	YES	464	130/176	73.86%	118/134	88.06%	216/240	90.00%		
Antidepressant	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
	NO	652	98/176	55.68%	60/134	44.78%	150/240	62.50%	344/362	95.03%
	YES	242	76/176	43.18%	74/134	55.22%	86/240	35.83%	6/362	1.66%
Mood Stabilizer	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
	NO	639	51/176	28.98%	57/134	42.54%	181/240	75.42%	350/362	96.69%
	YES	255	123/176	69.89%	77/134	57.46%	55/240	22.92%		
Anxiolytic	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
Sedatives	NO	741	120/176	68.18%	97/134	72.39%	181/240	75.42%	343/362	94.75%
Hypnotic	YES	153	54/176	30.68%	37/134	27.61%	55/240	22.92%	7/362	1.93%
Anticholinergic	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
Antiparkinsonian	NO	819	158/176	89.77%	114/134	85.07%	197/240	82.08%	350/362	96.69%
	YES	75	16/176	9.09%	20/134	14.93%	39/240	16.25%		
Miscellaneous,	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
Psychotropic	NO	880	170/176	96.59%	129/134	96.27%	231/240	96.25%	350/362	96.69%
Centrally Active	YES	14	4/176	2.27%	5/134	3.73%	5/240	2.08%		
Stimulants	Unknown	18	2/176	1.14%			4/240	1.67%	12/362	3.31%
	NO	855	156/176	88.64%	127/134	94.78%	223/240	92.92%	349/362	96.41%
	YES	39	18/176	10.23%	7/134	5.22%	13/240	5.42%	1/362	0.28%

Using the Alzheimer’s Disease Neuroimaging Initiative protocol (<http://adni.loni.usc.edu/methods/documents/mri-protocols/>), and appropriate for scanner brands, T1-weighted structural images were acquired on 3T magnets across different sites; T1-weighted magnetization prepared rapid acquisition gradient-echo or inversion

recovery-prepared spoiled gradient recoil sequences [53]. Table 3-4 shows the parameter settings used for the scanners.

Table 3-4 – B-SNIP dataset different parameter settings for the scanners

Scanner		shot interval	inversion time	TR	TE	flip angle	FOV	matrix	in-plane resolution	slices	slice thickness	voxel size	scan total duration
GE Signa	sagittal slab	2300 ms	700 ms	6.99 ms	2.85 ms	8 °	260×260 mm ²	256×256	1×1 mm ²	16	1.2m	1×1×1 0.2mm ³	10min 28sec
Philips Achieva	sagittal slab	3000 ms	846 ms	6.8 ms	3.1 ms	8 °	256×240 40 mm ²	256×240	1×1 mm ²	17	1.2m	1×1×1 0.2mm ³	9min 19sec
Siemens Allegra	sagittal slab	2300 ms	900 ms	7.2 ms	2.91 ms	9 °	256×240 40 mm ²	256×240	1×1 mm ²	16	1.2m	1×1×1 0.2mm ³	9min 14sec
Siemens Trio	sagittal slab	2300 ms	900 ms	6.8 ms	2.91 ms	9 °	256×240 40 mm ²	256×240	1×1 mm ²	16	1.2m	1×1×1 0.2mm ³	9min 14sec
GE Signa HDxt	sagittal slab	2300 ms	650 ms	7.0 ms	3.0 ms	8 °	256×256 56 mm ²	176x256x176	1×1 mm ²	16	1.2m	1×1×1 0.2mm ³	10min9sec
Siemens Trio	sagittal slab	2300 ms	900 ms	6.8 ms	2.74 ms	8 °	176×256 56 mm ²	176x256x176	1×1 mm ²	16	1.2m	1×1×1 0.2mm ³	10min9sec

Biotypes leverage neurobiological heterogeneity among psychosis cases in combination with information derived by laboratory tasks that assess brain function at the neurocognitive/perceptual level [52]. By integrating different biomarkers e.g. neuropsychological, stop signal, saccadic control, and auditory stimulation paradigms and other data external validating measures like social functioning, EEG, family biomarkers, and clinical information, nine variables were derived to capture neurobiological variance among the different psychosis cases which leads to three neurobiologically distinct psychosis biotypes which were distinctive and did not map to clinical boundaries. Note the structural MRI data were not used to derive the biotypes used in this study.

Probands and their first relatives were assessed with the structured clinical interview for DSM-IV to evaluate psychosis spectrum personality traits. Healthy subjects and their

first-degree relatives did not have history of psychotic or bipolar disorder records. The majority of probands and a minor subset of relatives were taking psychotropic medications [52].

For laboratory paradigms, there were multiple variables, five saccade variables two stop signal variables and 31 EEG variables. PCA data reduction applied for each paradigm set (saccades, stop signal task, EEG) which leads to two saccades, one stop signal, and five EEG components. Then using k-means clustering for class formation, 3 subgroups (biotypes) were used to capture cognitive-perceptual classification variance among the participants with psychosis [52].

3.2.2 Preprocessing

sMRI data (1-mm isotropic MPRAGE) were collected from all individuals and images were preprocessed in SPM (<https://www.fil.ion.ucl.ac.uk/spm>) via a unified approach [55] that included tissue classification, bias correction, image registration, and spatial normalization, and they were resliced to 2×2×2 mm. The unsmoothed gray matter density images were then correlated to the gray matter template to access segmentation outliers, and outliers were corrected if possible or removed otherwise. We calculated the correlation of each gray matter segmentation map to the mean map and removed those maps that had lower correlation (< 0.7). We have found these criteria to be very helpful in identifying problematic scans. We also centered and scaled the voxels of the subjects in each site to unit variance and covaried for each site separately.

3.2.3 Method Description

We used a novel classification/voting filtering method based on a data cleansing approach to eliminate label noise such as noisy mislabeled subjects based on the sMRI image dataset. This provided informative patterns for further investigation and for reassigning labels for identified mislabeled subjects based on the model. Our proposed model also provides suggested labels for noisy datasets. The method is based on computing m number of inner support vector machine (SVM) classifiers that are trained and evaluated via cross-validation. We used super vector classifiers from the scikit-learn Python library [56], and we used a one-versus-all approach that has been implemented in the scikit-learn library in the way that it fits one classifier for each class against all the other classes. These m SVM models are then used to identify mislabeled subjects for different runs of cross-validation sets. Thus, each individual is classified m times by SVM for k cross-validation loops totaling $m \times k$ classification votes. Based on the $m \times k$ predicted labels, we used consensus voting to determine whether a given dataset includes a noisy label if all $m \times k$ voting labels are mislabeled. Figure 3-1 presents a flow diagram of our method. In the following, we discuss more details regarding our proposed method.

3.2.3.1 Grid search

This approach seeks the hyper-parameter space through cross-validation and proposes the best candidate among possible hyper-parameters values.

1. Dimension reduction: We applied grid search consisted of 3 different dimension reduction algorithms, 1) selecting k feature using analysis of variance (ANOVA), which is computing a univariate ANOVA F-value between groups based on their assigned labels

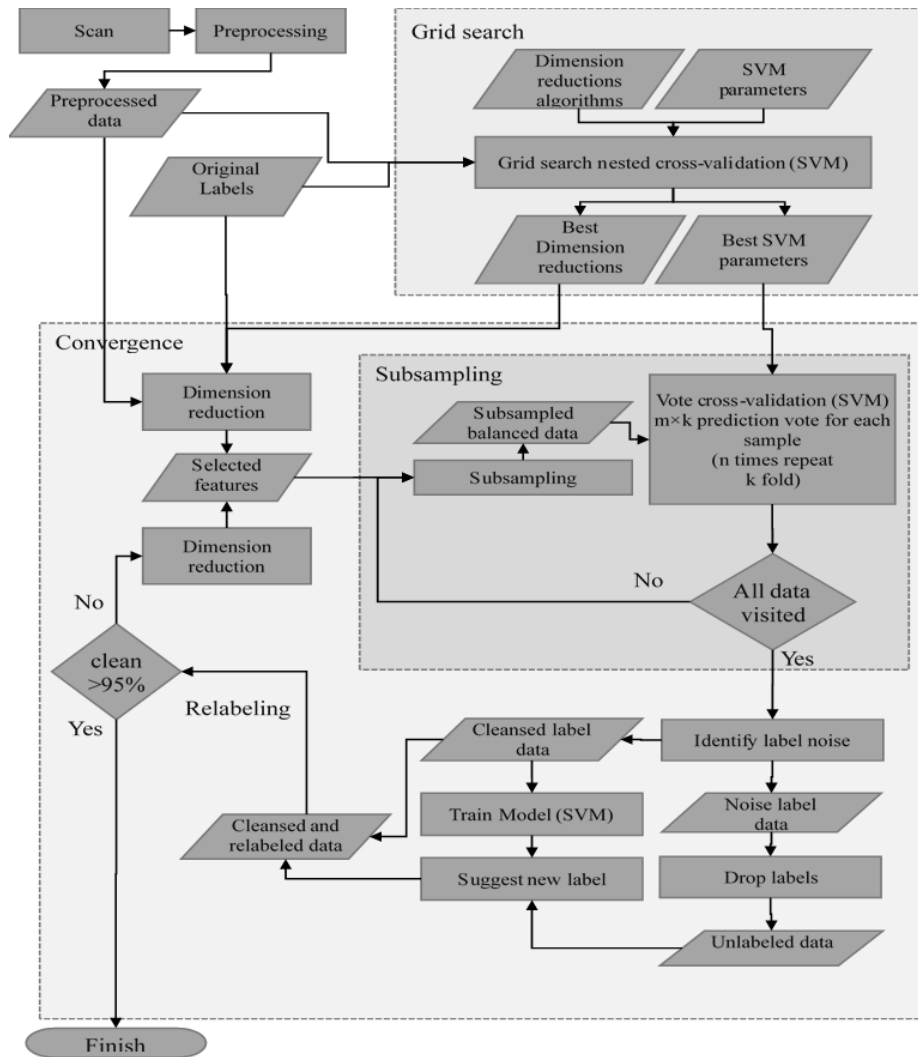


Figure 3-1 – Visual summary and workflow of different aspects of the proposed method. Preprocessing was done using SPM on structural MRI data collected at multiple sites. Hyper-parameters optimization and parameter selection among choosing the best dimension reduction were done using grid search. Then univariate feature selection using ANOVA was selected as best approach and used for the dimension reduction approach, applied for dimension reduction on preprocessed data and subsampling was done on dataset to handle imbalanced classes. Cross validated classification/voting filtering was applied to resampled data. Then noisy labels were identified based on their votes, cleansed dataset (exclude noise subjects) and unlabeled dataset (noisy subjects with dropped labels) feed to supervised to evaluate and obtain the new suggested labels. Combining cleansed data and data with new suggested labels make cleansed and relabeled data. For generalization, using a convergence approach, subsampling, cross-validated classification voting filtering and relabeling performed iteratively on cleansed and new labeled data till we identified a specific acceptable amount of label noise in the dataset.

and apply this to select the best features among vectors of features (vector of voxels) among instances. 2) Principal component analysis (PCA) which is a statistical procedure that uses an orthogonal transformation to find principal components which are uncorrelated to each other. 3) FastICA which is an efficient algorithm that performs independent component analysis (ICA). In this study we set 100 components for these three approaches.

2. Hyper-parameters optimization and parameter selection: Hyper-parameters are parameters within a learning algorithm which should be defined before model training and fitting. In contrast to ad-hoc hyper-parameters setting, hyper-parameter tuning/optimization refers to identifying the optimal set of hyper-parameters which maximizes the performance of the algorithm among the other options. In this work, to find the best hyper-parameters for the training model, we used a grid search among different hyper-parameters space and settings for the SVM model. Grid search for SVM classifier consists of super vector classifiers with 3 different nonlinear kernels (radial basis function, polynomial and sigmoid) as well as a linear kernel. The search was conducted over a kernel coefficient range of $(-4, 4)$ for the nonlinear kernels. Also $(0.0001, 1.0)$ was used as a regularization parameter space for grid search settings. We used nested cross validation by splitting data into two equal portions (50%) for training and testing in inner cross validation and using 5-fold cross validation for outer cross validation.

3.2.3.2 Dimension reduction and feature extraction

The curse-of-dimensionality comes with high dimensional neuroimaging scan data where the number of features (voxels) is significantly larger than the number of samples (brain images) [58]. This is problematic because it may lead to overfitting in the predictive

models and a lack of generalizability. We used univariate feature selection and computed a univariate ANOVA F-value between groups based on their assigned labels and used this to select the 100 best feature vectors (a vector of voxels) for each cleansing iteration.

3.2.3.3 Imbalanced classes and subsampling

A class imbalance problem occurs when distributions of different classes are imbalanced, and each class does not represent an equal portion of the dataset. Imbalanced classes appear in most real-world datasets. Depending on the classification task, an imbalanced dataset can impact the accuracy of the classification [63]. This is because many classifiers assume the same distribution and proportion of samples across classes. More samples from a specific class will result in a biased result in favor of those classes [64]. We therefore used a random under-sampling approach for handling imbalanced classes by resampling the majority class randomly and uniformly. This was done repeatedly until all the instances of the majority classes were visited at least one time.

3.2.3.4 Data cleansing classification/voting filtering

We obtain votes using cross-validated classification. Using the consensus voting approach, we considered individuals as noisy if all votes are inaccurate labels. Based on this, we keep all the data but remove the labels for noisy subjects and consider them to be an unlabeled sample for relabeling steps.

3.2.3.5 Relabeling using classification

After removing labels from the noisy label data, we trained the SVM model with cleansed data and predicted a new label for noisy subjects. The classification relabeling

approach assumes that when two samples in a high-density region are close and similar then their output classification should be close too.

3.2.3.6 Convergence

After relabeling step, we repeated the process from subsampling step to relabeling step iteratively with new suggested labeled till the number of label noise identified in filtering step reach to an ad-hoc threshold (e.g. 5%). In each iteration, subsampling, classification voting filtering and relabeling perform on the cleansed and relabeled datasets (i.e. the entire dataset with original labels for cleansed data and newly assigned labels for label noise subjects). By using this iterative step, label noises gradually are detected and the performance of the model increases. Combining this iteration and randomly chosen partition for training and validating increase generalization of our approach.

3.2.3.7 Visualization

After our iterative classification filtering, we performed voxelwise t-tests on gray matter maps to highlight brain regions where different groups differed significantly. For each pair of groups, we ran statistical t-tests on the original non-zero vectors of features (voxels) and then used the false discovery rate (FDR) correction for multiple comparisons. We identified those features that were significantly different between each pair of groups and created images only containing means of larger group of those features. We plotted the dataset before and after cleansing on a 2D plot using 2D-tSTE [65].

3.2.3.8 Simulation

Because there is no ground truth in the neuroimaging data, we first evaluated our method on a handwritten digit images dataset, which enabled us to compare our method on data for which there is well-defined ground truth. We introduced label noise to the dataset by shuffling a proportion of labels of instances randomly. We added different amounts of label noise and evaluated the model. We included images of digits 1, 3, 5, 8 because their shapes' similarity makes the prediction more challenging. The images have dimension 64 pixels (8×8) Figure 3-2. We selected and shuffled truth labels of randomly chosen digit images of different proportions in the dataset.

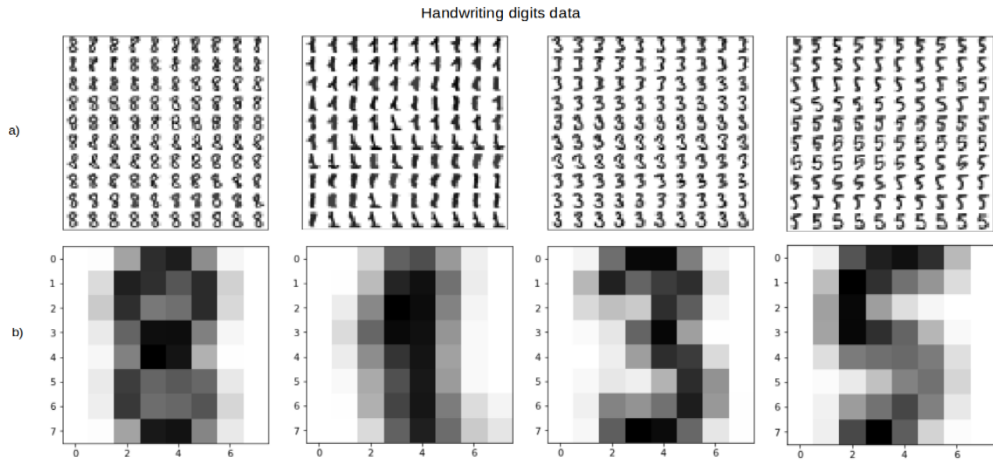


Figure 3-2 – a) Shows random samples from handwriting digits dataset. The numbers of 8,1,3,5 were selected because of their similarities and common features they have; and b) The mean over all samples for each digit.

One potential concern that may arise is that we do not have a ground truth to assess if the model is biased. In the presence of label noise, the validation and test accuracy of the model can be decreased by noisy subjects, and likewise it is not possible to estimate generalization metrics without a ground truth [38]. We appeal to the simulation as a proxy for validating our model as it is not possible to know the ground truth in real data. We analyzed our method on the simulated handwriting digit dataset where identification of the

label noise is possible since we have the ground truth. We artificially introduced label noise into the simulation dataset by shuffling their labels. Results showed our method successfully identified the correct labels and the performance of the model increased significantly on new cleansed data.

3.2.3.9 Deep learning

We also evaluated our relabeling dataset using another model, a deep learning residual model, to elucidate the effect of label noise on the accuracy of the predictive model by comparing accuracy before and after a data cleansing classification filtering method. We analyzed the cleansed dataset with a deep convolutional neural network (CNN) ResNet model. Using a deep learning model, we can circumvent feature engineering which is basically extracting useful representations and compare its result without incorporating an explicit dimension reduction step. The architecture of our model is the modification of the open-source Pytorch implementation of the ResNet framework. 3D structural MRI scans are seen by the convolution layer followed by batch normalization and a rectifier unit layer. The output of these layers goes through the max pooling layer and then goes through the Residual block of containing a series of 3D convolutional, batch normalization and non-linear rectifier (ReLU) layers [46]. The output of residual blocks feeds to into the average pooling layer followed by fully connected layers. The output of this layer is the class probability score. See Figure 3-3.

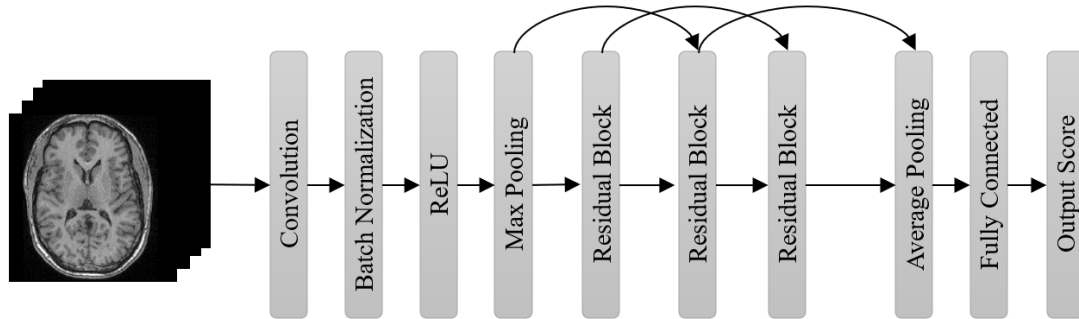


Figure 3-3 – Deep model convolutional neural network ResNet architecture. 3D structural MRI scans are seen by the convolution layer followed by batch normalization and a rectifier unit layer. The output of these layers goes through max pooling layer and then go through Residual block of containing a series of 3D convolutional, batch normalization and non-linear rectifier (ReLU) layers

3.3 Results

3.3.1.1 Simulation

Figure 3-4 shows that by increasing the amount of label noise, the accuracy of the SVM classifier model decrease linearly on original dataset; however, by applying our method we obtain high accuracy on the cleansed dataset by identifying noise label instances and then filtering them out from the dataset. It is also important to note that having label noise in the dataset affects the evaluation part for unforeseen data. However, considering grand truth in simulation shows label noises on unseen data can be detected and correct labels assigned to them with high accuracy.

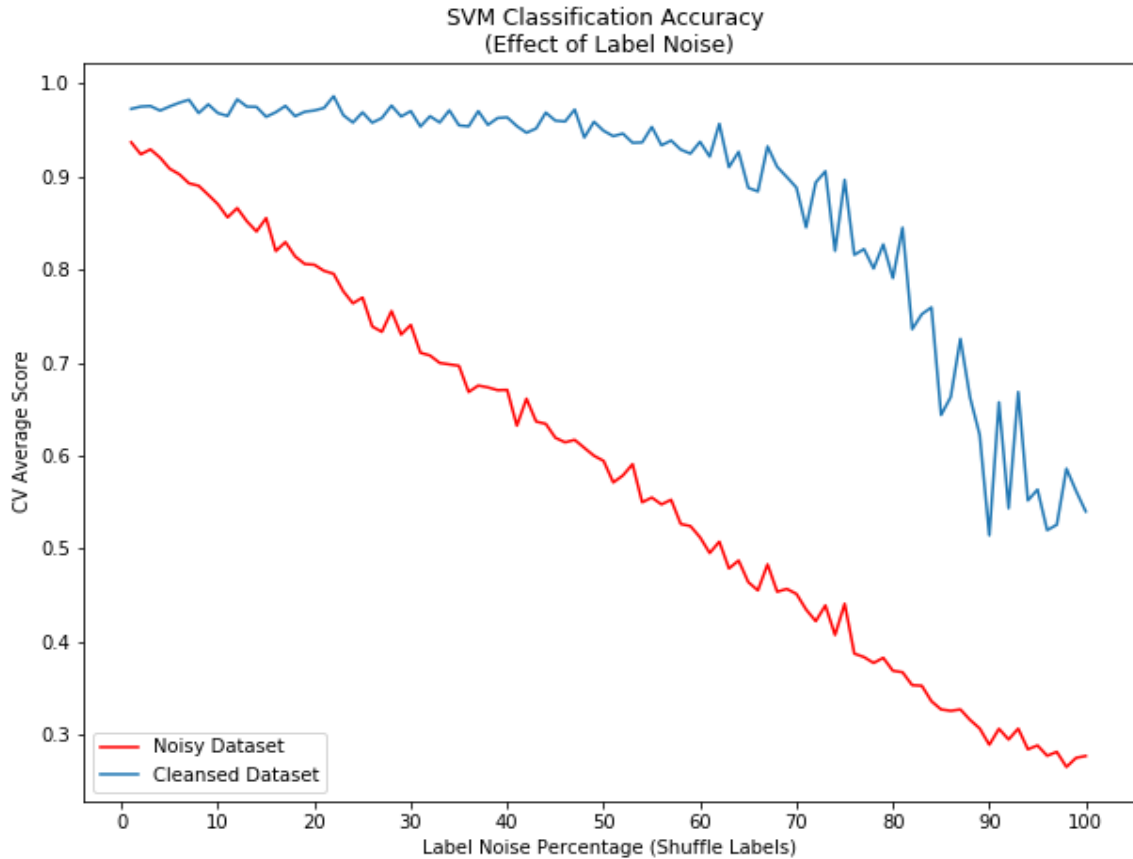


Figure 3-4 – Shows effect of different proportion of label noise on accuracy of SVM classifier and improving accuracy using data cleansing approach. Corrupting data by shuffling labels of instances and adding synthetic label noises to ground truth decreases the performance of the predictive model. By applying data cleansing classification filtering method, we could identify noisy labels and improve the model performance. Even when 80% of labels in dataset are noisy, model accuracy was boosted from 0.38 in the noisy dataset to 0.9 in the cleansed dataset. The effect of label noise on splitting dataset also shows evaluating data cleansins is challenging when the ground truth are not available and may mislead during interpretation.

3.3.1.2 B-SNIP MRI Analysis

After validating our approach on simulation data, we applied our method on the B-SNIP dataset. For the B-SNIP data, we used a univariate ANOVA F-value to select the 100 best features among 2,122,945 voxels for 912 individuals. Features extracted were most often found in amygdala, cerebellum and insula regions bilaterally. Some additional

features which recurred included lingual, occipital, frontal, temporal, fusiform, para hippocampal and hippocampus regions. This was done for each type of labelling approach (Biotype and DSM). To handle imbalanced classes, we used the subsampling approach described previously.

Using a grid search approach, the SVM classifier with RBF kernel and penalty parameter 10 and kernel coefficient 0.04 was selected as the best model with optimized hyper-parameters for the B-SNIP dataset. The average accuracy of classification filtering is shown in Figure 3-5. The average SVM classifier cross validated accuracy increased from 0.38 to 0.89 for both Biotype and DSM-IV categories using consensus voting.

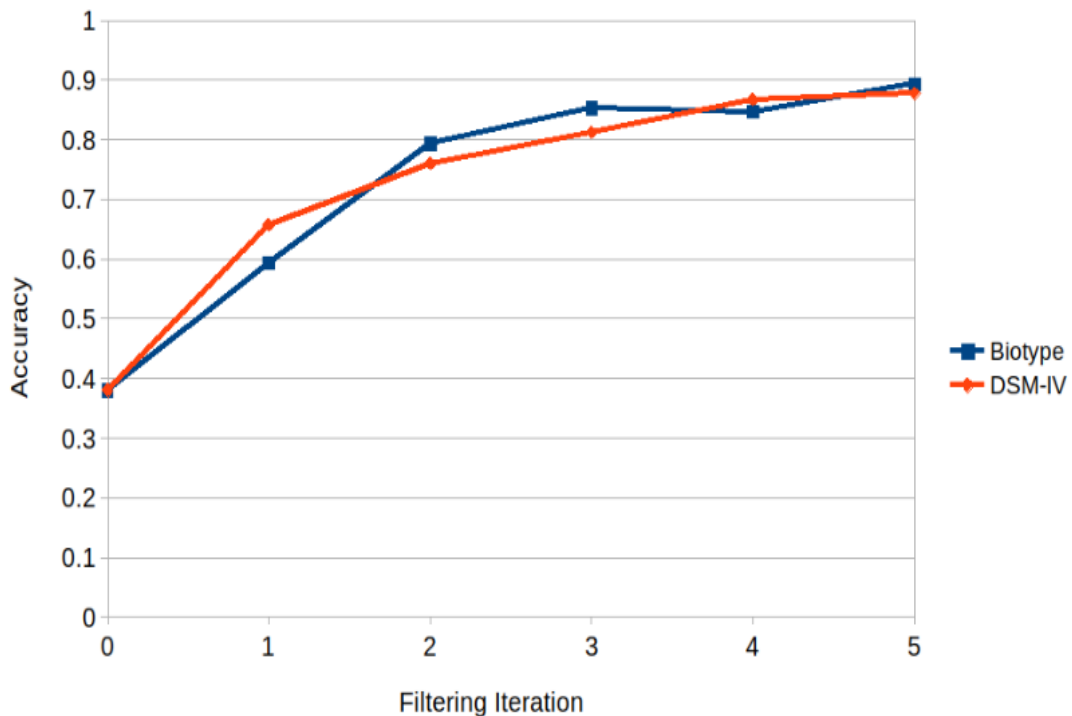


Figure 3-5 – Average accuracy of classification filtering cross validation part after 5 iterations. In each iteration, label noise is identified, and new labels are assigned based on similarities between them and cleansed instances. The performance of the model increases after each iteration.

Among 912 subjects, 573 subjects (63%) were identified as label noise in the Biotype category and 601 subjects (65%) were identified as label noise in the DSM-IV category. 452 out of 912 labels were identified as noisy for both Biotype and DSM-IV. Table 3-5 shows the different proportion and number of noisy labels found after 5 iterations of cross validation classification filtering and reveals convergence at approximately 90% of data has been cleansed. At each iteration we removed the labels for the noisy datasets (retaining the labels for the clean datasets) and fit an SVM model to predict new suggested labels. Figure 3-6 shows the heatmaps of those identified label noise in different types of labels. Also using the fit model, we predict the label for the relatives. 435 out of 581 relatives are labeled as Biotype B3 and normal control in the Biotype category and all of 581 are labeled as controls in the DSM-IV category.

Table 3-5 – Shared label noise using consensus voting

Consensus Biotype								Consensus DSM-IV							
Clean:460				Noise: 452				Clean: 460				Noise: 452			
B1		B2		B3		NC		BPD		SADP		SZP		HC	
Clea	Nois	Clea	Nois	Clea	Nois	Clea	Nois	Clea	Nois	Clea	Nois	Clea	Nois	Clea	Nois
n	e	n	e	n	e	n	e	n	e	n	e	n	e	n	e
63	84	78	107	101	117	218	144	89	87	69	65	84	156	218	144

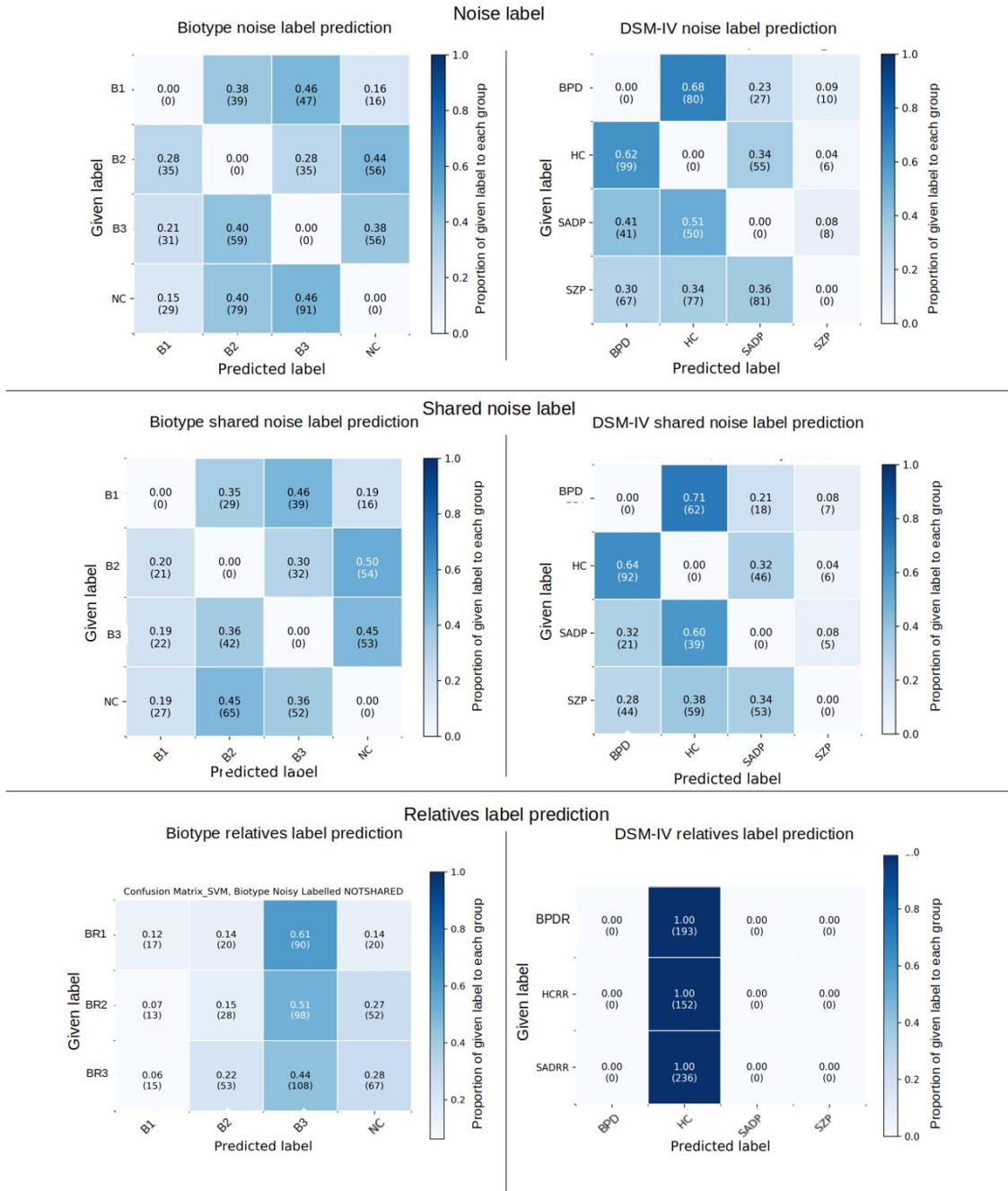


Figure 3-6 – Top: Confusion matrices show label noise subjects related to Biotype or DSM-IV categories separately after convergence to 90% of cleansed data using consensus voting. Middle: Confusion matrices show shared label noise subjects among Biotype and DSM-IV categories after convergence. About 50% of Biotype proband 2 (B2) and 45% of Biotype proband 3 (B3) were relabeled as normal control. Among Biotype proband 1 (B1), 81% of instances identified as noisy label were relabeled as Biotype proband 2 (B2) and Biotype proband 3 (B3). 81% of normal control noisy subjects were also relabeled as Biotype proband 2 (B2) and Biotype

proband 3 (B3). In DSM-IV category, 71% of bipolar proband noisy subjects were relabeled as healthy controls, whereas 64% and 32 % of healthy controls were relabeled as bipolar proband and schizoaffective proband respectively. Among schizophrenia label noise subjects, 38% were relabeled as healthy controls and 34% as schizoaffective proband. Bottom: Confusion matrices show the result of predicting labels of relatives using updated labels after the above analysis. Most of the relatives are labeled as Biotype3 (B3) and normal controls in the Biotype categories or as healthy controls in the DSM-IV category.

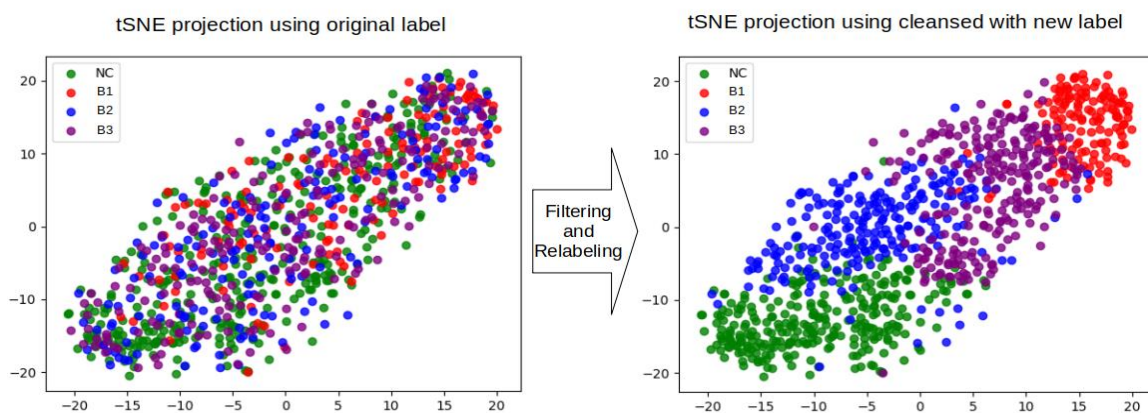
Among the evaluated criteria, the label noise found in DSM-IV showed more irregularity than Biotype criteria after iterative classification filtering. Among 912 subjects in different groups of DSM-IV, we found that the schizophrenia group showed the most differences. By comparing given labels and new assigned labels, the proportion of schizophrenia group reduced from 26% to 4% of total subjects after convergence and only 6% of schizophrenia subjects remained in the same category after filtering and relabeling. About 62% of schizophrenia subjects were assigned new labels in schizoaffective and bipolar disorder categories and 32% of them were relabeled as healthy controls. Also, from 912 subjects, about 6% of the bipolar group, 2% of healthy control and 6% of the schizoaffective group, were relabeled as schizophrenia. Among different groups of patients in DSM-IV, 50% of bipolar, 20% of schizoaffective and 42% of schizophrenia control were relabeled as healthy control. 56% of healthy controls remained in their group and from remaining healthy control noisy subjects, 27% of them categorized as bipolar, 15% relabeled as schizoaffective and 2% were relabeled as schizophrenia.

On the other hand, for the Biotype groups, about 32% of subjects remained in their original Biotype patient groups and 45% of normal controls remained in their group. Normal controls and Biotype category 3 (B3) changed subjects between each other more than any other two groups in the Biotype criterion. This may be because Biotype categories

are more based on neuromarkers than clinical symptoms and Biotype category 3 is biologically most similar to normal controls [61]. For the Biotype cleansed relabeled dataset, 51% of relatives were categorized as Biotype B3, 24% of them were labeled as normal control, 17% of them were categorized as Biotype B2 and 8% were categorized as Biotype B1.

We used t-SNE 2D-projection of the dataset using the original and suggested new labels to visualize how well subjects will be categorized based on their labels. First, 100 features were extracted from nonzero voxels using the univariate feature selection method. Next, these features were projected in the 2D plot using tSNE. Figure 3-7 shows the t-SNE 2D-projection of the original dataset with label noise in the left panel, and projection using the new suggested labels in the right panel.

Biotype tSNE 2D-Projection



DSM-IV tSNE 2D-Projection

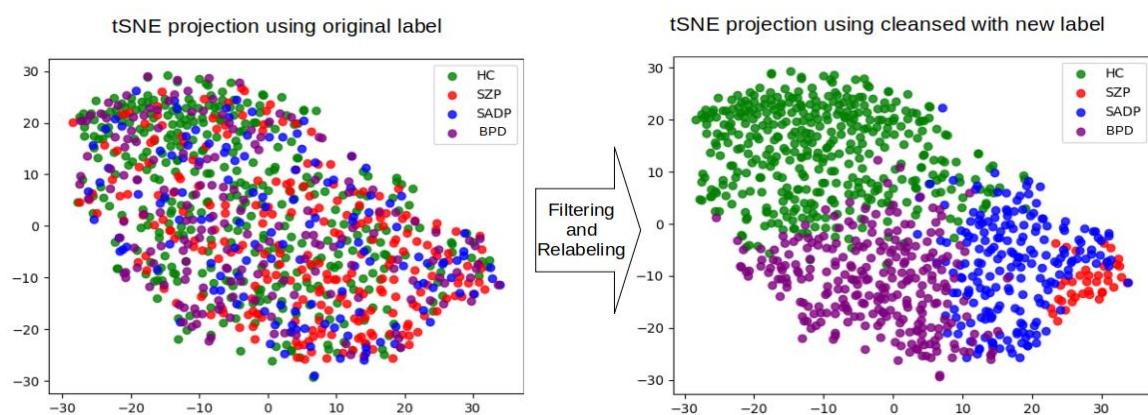


Figure 3-7 – Left panel shows t-SNE 2D-projection of original dataset with label noise and right panel shows 2D-projection using new suggested labels. Top: 2D-projection using Biotype labels. Bottom: 2D-projection using DSM-IV labels. In both categories of Biotype and DSM-IV, affinities between data points shows subjects in different group of disease and healthy control overlap in noisy data. After identifying label noise and relabeling, the similarity between data points is more obvious and subjects with the same labels are close together. Embedded data into 2D space using original labels does not support the fact that subjects labeled based on their similarities. It is hard to interpret how groups differentiate from each other using original label, because there is considerable overlap between subjects. Cleansed dataset with new suggested label shows there is gradient pattern in both DSM-IV and Biotype categories from healthy/normal control to most severe cases of Biotype B1 or schizophrenia probands contain other mild cases in between.

Voxel-wise t-tests were run on non-zero voxels between each category. Then FDR correction was applied to correct for multiple comparisons. Voxels showing significant

differences at a significance level $\alpha=0.05$ were identified. Figure 3-8, Figure 3-9, Figure 3-10, and Figure 3-11 show the results of statistical testing between each group. Figure 3-8 and Figure 3-9 show gray matter brain maps results using Biotype cleansed labels and Biotype original given label (old label) containing noise. Figure 3-10 and Figure 3-11 show gray matter brain maps results using DSM-IV cleansed labels and DSM-IV original diagnostic labels. Results who more voxels showing significant differences when using the cleansed data, and more voxels showing significant differences in the DSM compared to the Biotype data. Importantly the regions that are shown are consistent with previous work [12], [47] (for example in schizophrenia we find reductions in bilateral temporal and insula regions as well as medial frontal).

Table 3-6 shows regions where the statistical tests are significantly different after FDR multiple comparisons correction on the p-values. When using the original dataset, we did not find differences in as many regions. There were significant differences in some regions between healthy controls and schizophrenia and bipolar probands and schizophrenia. Significantly different brain regions for the relabeled data are indicated with ● and by ◇ in the original dataset.

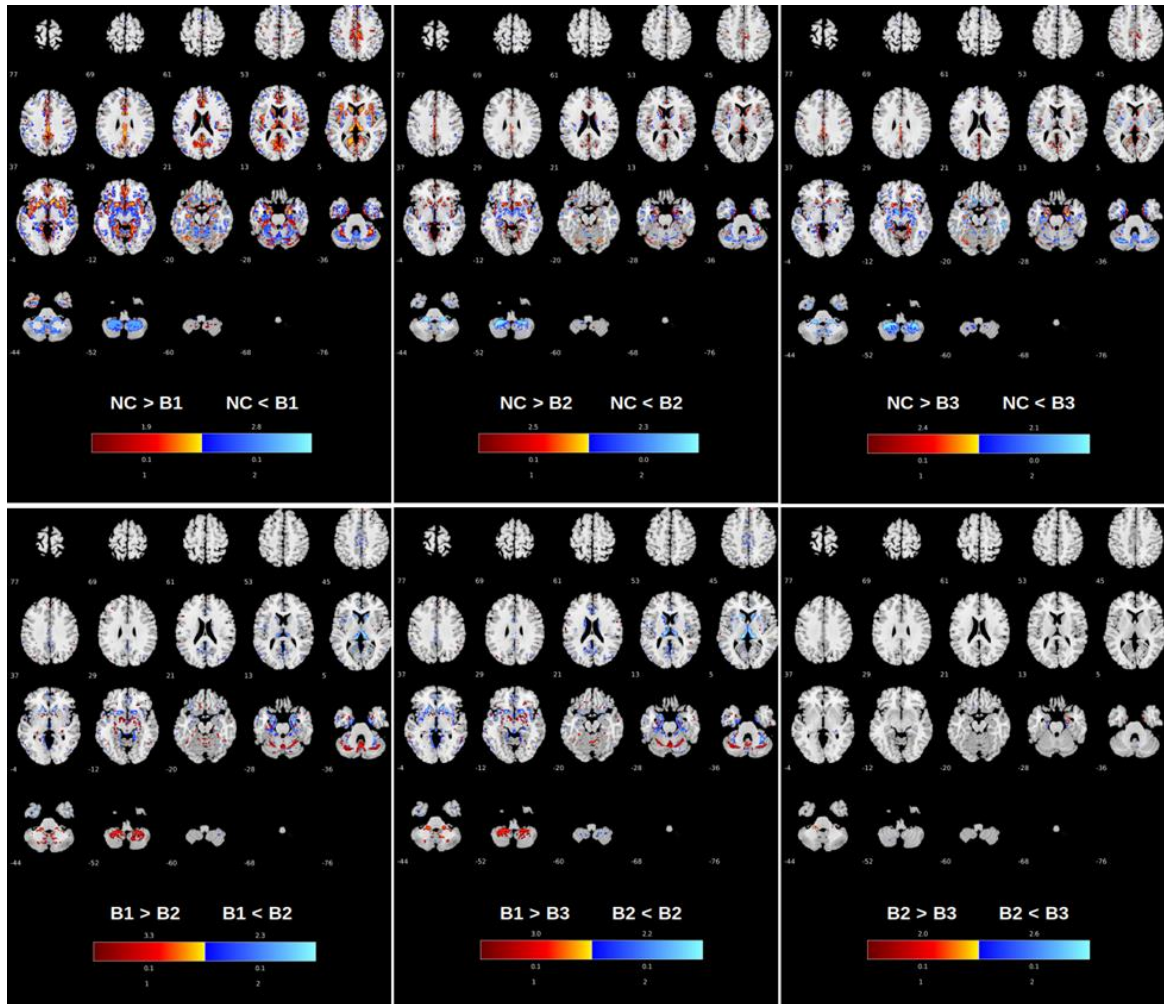


Figure 3-8 – Gray matter map results of voxel wise t-tests between 4 Biotype groups after data cleansing using classification voting filtering. Top row) NC vs. B1, NC vs. B2, NC vs. B3. Bottom row) B1 vs. B2, B1 vs. B3, B2 vs. B3. Gray matter contrast between normal controls and Biotype probands shows they have differences with different levels in overlapped regions after cleaning and relabeling individuals. Gray matter differences between normal control and Biotype proband 1 (B1) have strongest separation among the other comparison tests. The gray matter contrast is lower between healthy controls and Biotype proband 3 (B3). Biotype proband 2 (B2) and Biotype proband 3 (B3) show the fewest differences.

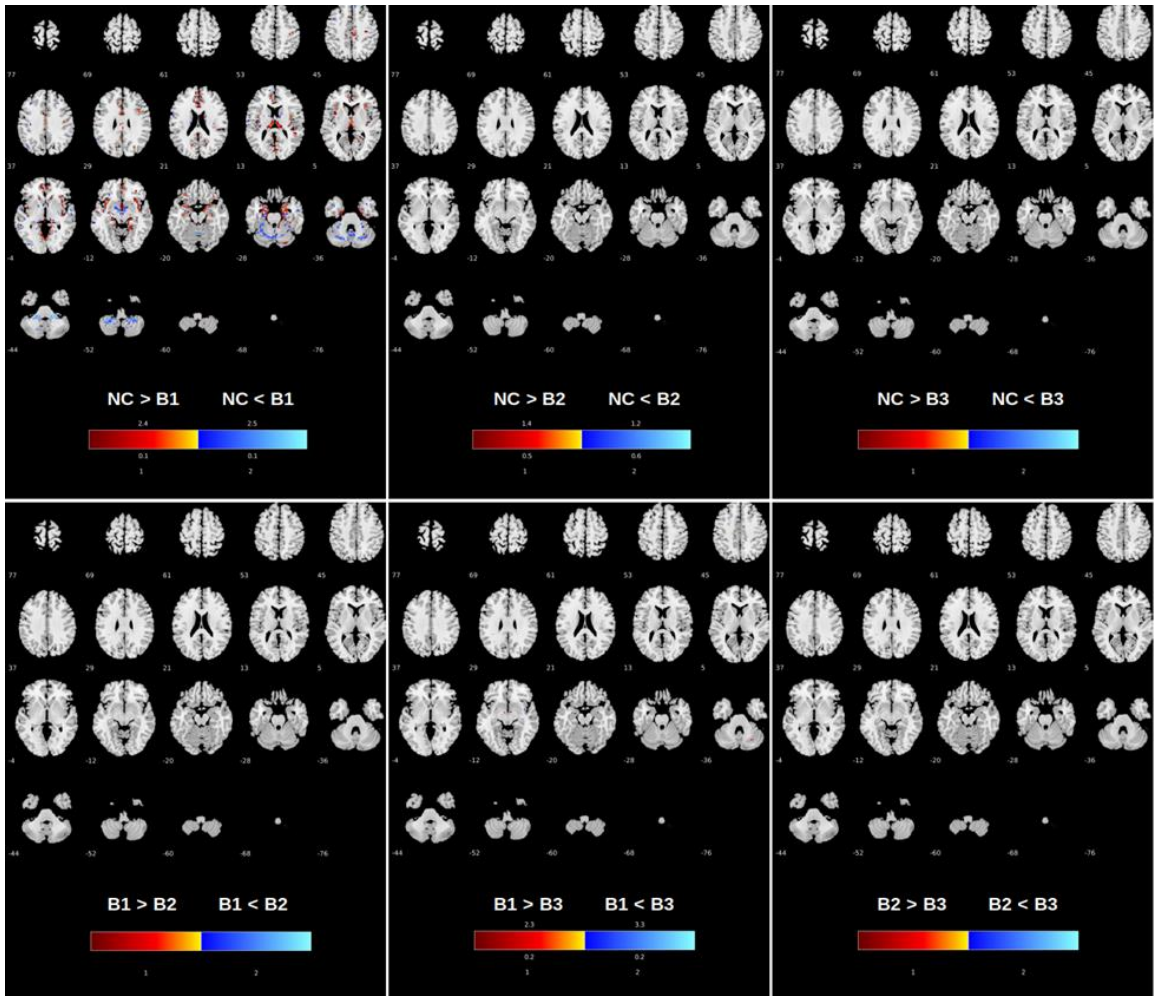


Figure 3-9 – Gray matter map results of voxel wise t-tests between 4 Biotype groups on given labels. Top row: NC vs. B1, NC vs. B2, NC vs. B3; Bottom row: B1 vs. B2, B1 vs. B2, B2 vs. B3. Gray matter contrast between normal controls and Biotype probands only shows significant difference in some regions between normal controls and Biotype proband 1 (B1) using original labels. Gray matter density group differences between other groups do not show differences.

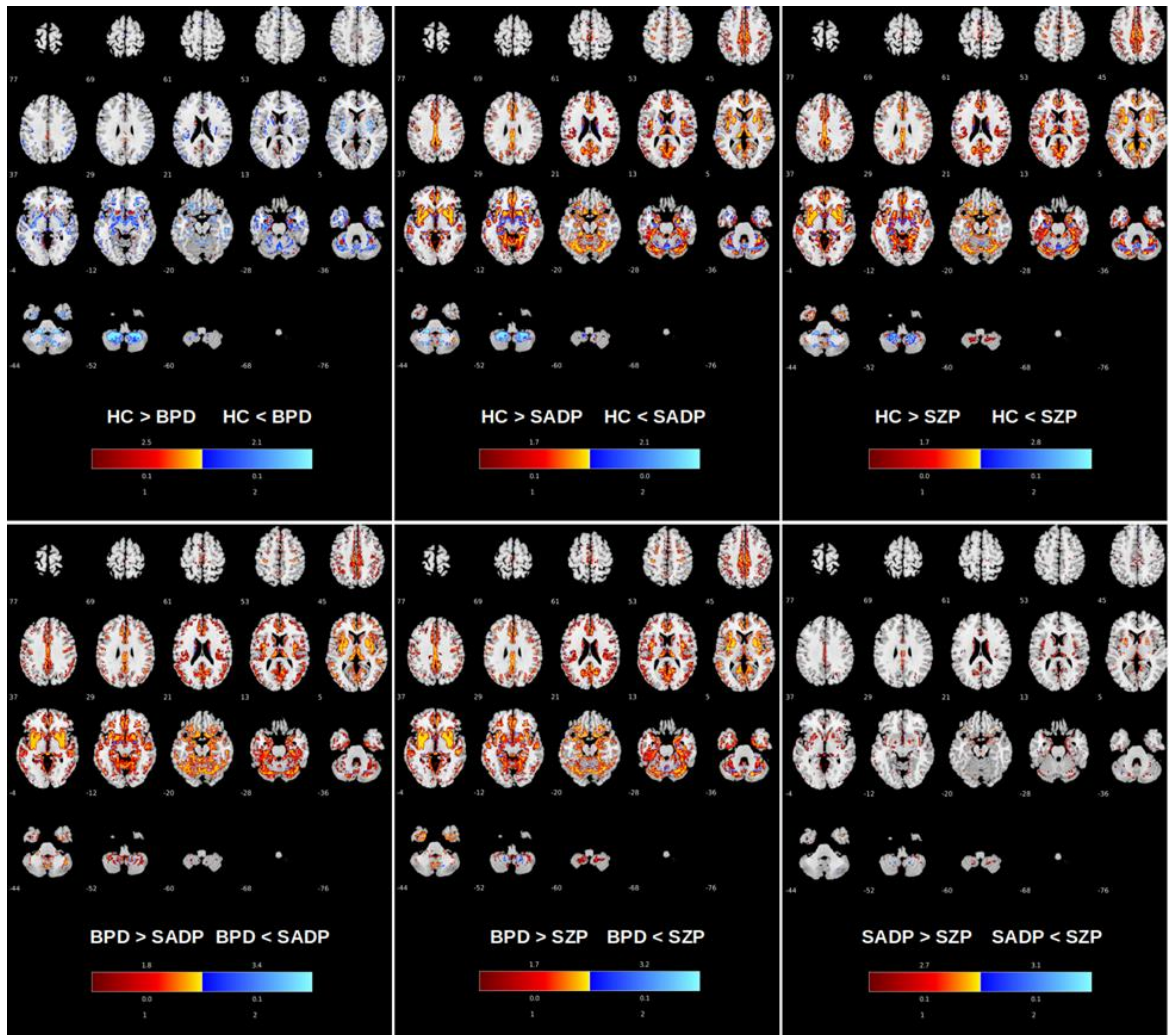


Figure 3-10 – Gray matter map results of voxel wise t-tests between 4 DSM-IV groups after data cleansing using classification voting filtering. Top row) BPD vs. HC, HC vs. SADP, HC vs. SZP. Bottom row) BPD vs. SADP, BPD vs. SZP, SADP vs. SZP. Gray matter contrast between healthy controls and DSM-IV probands shows they have differences with different levels in some overlapped regions after cleaning and relabeling individuals. Gray matter difference between healthy control and bipolar proband (BPD) has strongest separation among the other. The gray matter contrast shows differences between healthy control and schizophrenia and schizoaffective probands as well. Also, group differences could be observed between the bipolar proband group vs. schizophrenia and bipolar proband group vs. schizoaffective proband groups. However, schizophrenia and schizoaffective groups did not separate after data cleansing and relabeling.

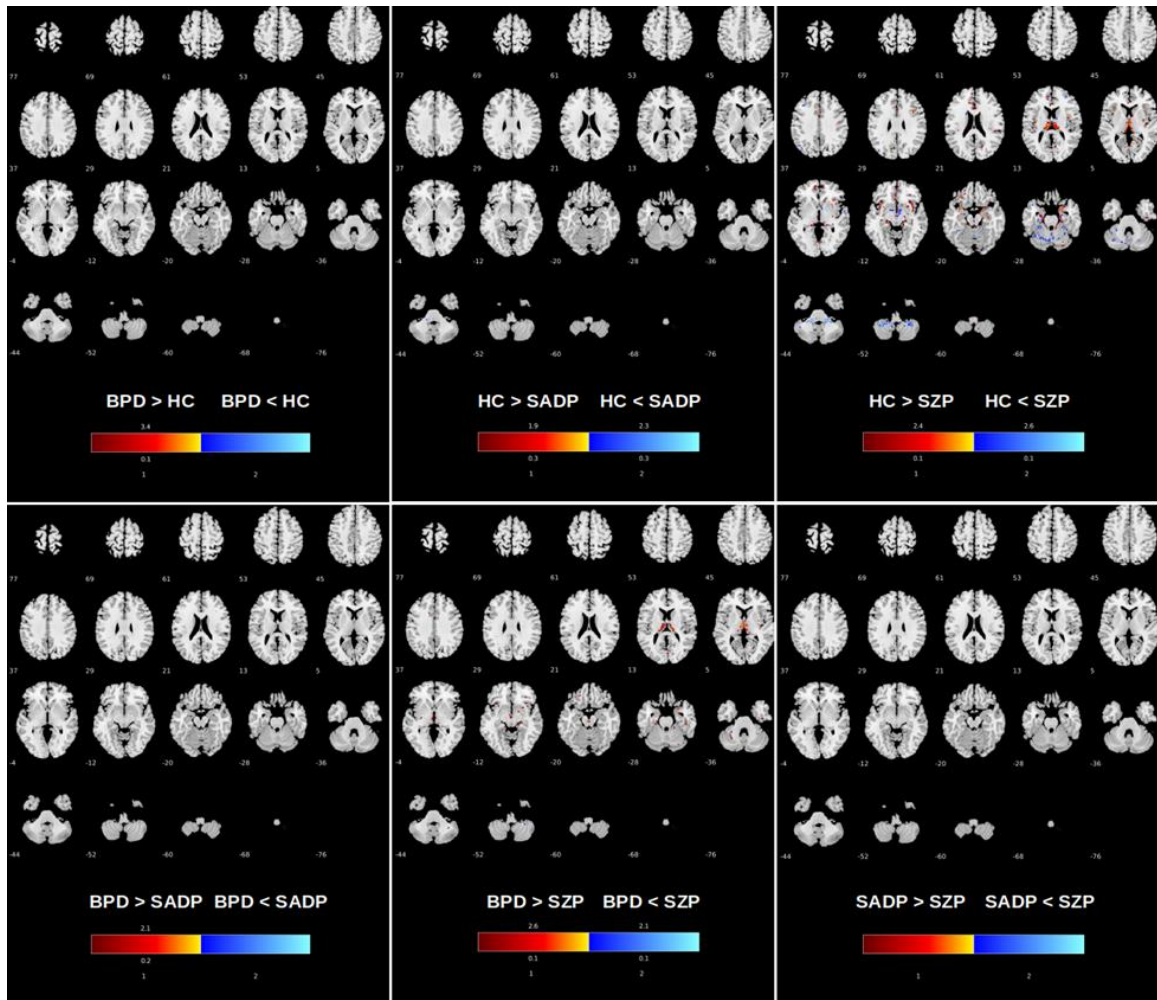


Figure 3-11 – Gray matter map results of voxel wise t-tests between 4 DSM-IV groups on given labels. Top row: BPD vs. HC, HC vs. SADP, HC vs. SZP. Bottom row) BPD vs. SADP, BPD vs. SZP, SADP vs. SZP. Gray matter contrast was only significant in some areas between healthy controls and schizophrenia proband using the original labels. Group differences between other groups were not found.

Table 3-6 – Area of containing voxels obtained by statistical significant test on the relabeled data ● and original data ◇

<i>Area</i>	B1 Vs. B2	B1 Vs. B3	B2 Vs. B3	NC Vs. B1	NC Vs. B2	NC Vs. B3	BDP Vs. HC	BDP Vs. SADP	BDP Vs. SZP	HC Vs. SADP	HC Vs. SZP	SADP Vs. SZP
<i>Angular Gyrus</i>								●	●	●	●	
<i>Anterior Cingulate</i>		●		●◇		●	●	●	●	●	●◇	
<i>Caudate</i>											●	●
<i>Cerebellar Lingual</i>										●		
<i>Cerebellar Tonsil</i>	●	●◇	●	●◇	●	●	●	●	●	●	●◇	●
<i>Cingulate Gyrus</i>				●◇				●	●	●◇	●◇	
<i>Clastrum</i>								●				
<i>Culmen</i>	●	●		●◇	●	●	●	●	●	●◇	●◇	●
<i>Cuneus</i>				●◇	●	●	●	●	●	●	●◇	
<i>Declive</i>				●◇		●	●	●	●	●	●◇	
<i>Extra-Nuclear</i>	●	●		●◇	●	●	●	●	●	●◇	●◇	●
<i>Fusiform Gyrus</i>				●		●	●	●	●	●◇	●	
<i>Inferior Frontal Gyrus</i>	●	●◇		●◇	●	●	●	●	●	●	●◇	●
<i>Inferior Occipital Gyrus</i>				●				●	●		●◇	
<i>Inferior Parietal Lobule</i>				●◇				●	●	●	●	
<i>Inferior Temporal Gyrus</i>				●	●	●	●	●	●	●	●	
<i>Insula</i>	●	●		●◇			●	●	●	●◇	●◇	●
<i>Lateral Ventricle</i>				◇	●		●			●	●	
<i>Lentiform Nucleus</i>				◇	●	●	●	●	●	●◇	●◇	
<i>Lingual Gyrus</i>				●◇	●	●	●	●	●	●	●	
<i>Medial Frontal Gyrus</i>	●	●		●◇	●		●	●	●	●	●◇	
<i>Middle Frontal Gyrus</i>				●◇		●	●	●	●	●	●	
<i>Middle Occipital Gyrus</i>				●◇				●	●	●◇	●◇	
<i>Middle Temporal Gyrus</i>				●◇	●		●	●	●	●	●◇	
<i>Nodule</i>				◇	●◇	●	●◇	●			●◇	
<i>Paracentral Lobule</i>	●			●◇	●	●	●	●	●	●	●	
<i>Parahippocampal Gyrus</i>	●	●	●	●◇	●	●	●	●	●	●◇	●◇	●
<i>Postcentral Gyrus</i>				●◇				●	●	●	●◇	
<i>Posterior Cingulate</i>				●				●	●	●	●	●
<i>Precentral Gyrus</i>	●	●		●◇			●	●	●	●	●	
<i>Precuneus</i>				●◇	●		●	●	●	●◇	●◇	
<i>Pyramis</i>	●	●		●◇	●	●	●	●	●	●	●	●
<i>Sub-Gyral</i>	●	●		●◇	●	●	●	●	●	●◇	●◇	
<i>Subcallosal Gyrus</i>	●	●		●	●	●	●	●	●	●	●	
<i>Superior Frontal Gyrus</i>				●◇	●	●	●	●	●	●	●◇	
<i>Superior Parietal Lobule</i>								●	●	●	●	
<i>Superior Temporal Gyrus</i>		●		●◇	●	●	●	●	●	●	●◇	
<i>Supramarginal Gyrus</i>								●	●	●	●	
<i>Thalamus</i>		●		●◇			●	●◇	●	●◇	●◇	●
<i>Transverse Temporal Gyrus</i>				●		●	●	●	●	●	●	
<i>Tuber</i>				●◇	●	●	●	●	●	●	●	
<i>Uncus</i>		●◇		●	●	●	●	●	●◇	●◇	●◇	
<i>Uvula</i>				●◇	●	●	●	●	●	●	◇	

3.3.1.3 Deep learning model

We analyzed the performance of our deep learning model on the first iteration of classification filtering. The accuracy of our deep supervised convolutional neural network (ResNet) model improved about 20% from 0.65 to 0.79 on an average of 6 binary classification task using DSM-IV labels and by 23% from 0.60 to 0.74 on 6 binary classification task using Biotype labels. Figure 3-12 shows the receiver operating characteristic (ROC) plot of these 12 binary classification tasks on our deep model. The boxplot of the accuracy of 12 classification tasks on cleansed and original dataset shows in Figure 3-13.

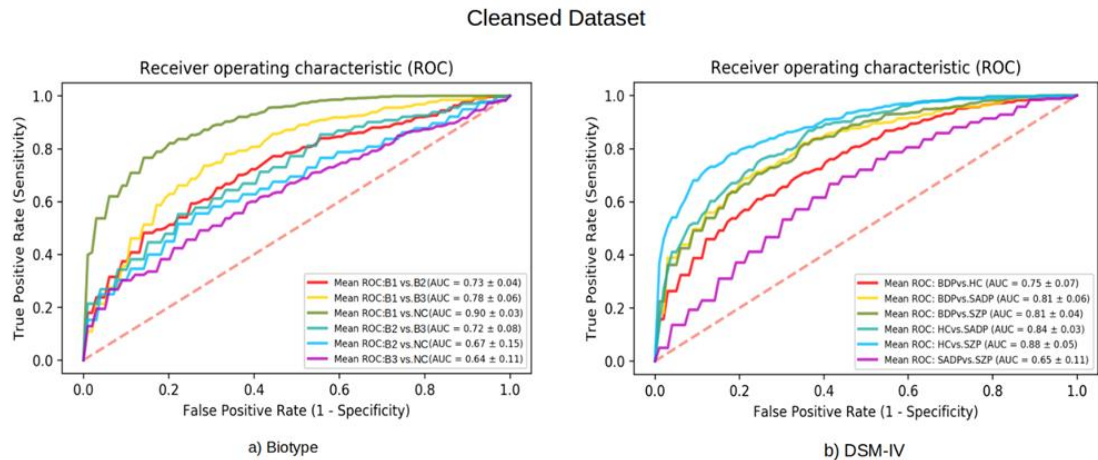


Figure 3-12 – ROC plot of 12 binary classification tasks using Biotype and DSM-IV labels on cleansed dataset after 1st iteration of our data cleansing approach before convergence. After one iteration of classification filtering on data and training deep model using cleansed dataset the accuracy improved about 20% from 0.65 to 0.79 on an average of 6 binary classification task using DSM-IV labels and by 23% from 0.60 to 0.74 on 6 binary classification task using Biotype labels.

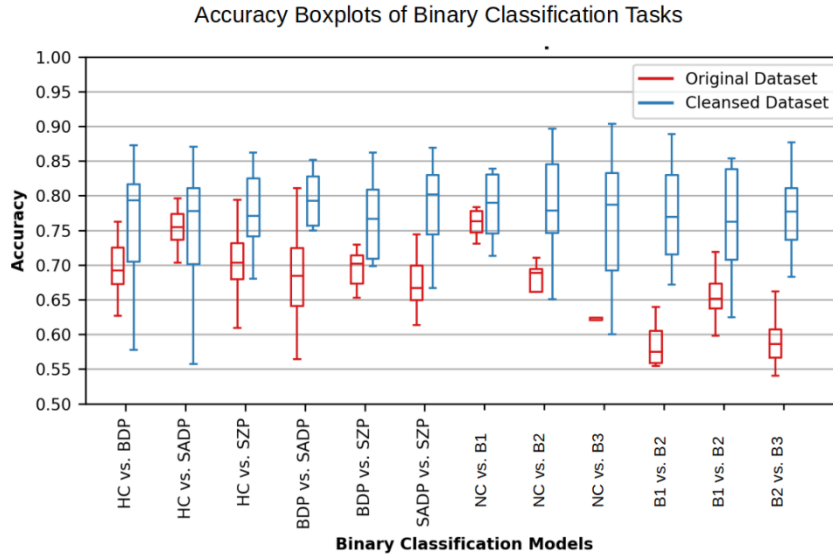


Figure 3-13 – Accuracy boxplot of 12 binary classification tasks on deep model on cleansed dataset after 1st iteration. Red are boxplot accuracy of stratified cross validation using original label. Blue boxplot shows the accuracy of different stratified cross validation run on cleansed dataset after a single classification voting filtering. The figure shows accuracy improved on cleansed dataset after a single iteration of classification filtering.

3.4 Discussion

Current diagnostic categories for mental illness are not biologically based and exhibit considerable heterogeneity within diagnosis and overlap across diagnoses [66], [67]. Most imaging biomarker studies do not surprisingly use the DSM categories as the ground truth, which, while informative, does not help us move beyond the known validity issues. In this work, we present a first towards addressing the known problems with existing labeling schemes but providing a way for them to be updated by biological data (e.g. brain imaging). The approach we use uses a current categorization approach as input but then updates are using additional biological data by assuming there is noise in the assignment process. We show that the MRI data is pushing against the labeling assigned via two different diagnostic

categorization approaches (DSM & biotypes). Then using existing categories, we ask how the data should be categorized within each scheme to reduce the label noise. We are currently exploring extending this work via semi-supervised approaches and to nominate new categories based on the biological data alone. This is an extremely difficult task which has been underexplored in the context of brain imaging data and we hope to contribute more on this going forward.

Many studies have been done to address label noise issues in different domains such as medical imaging [68]–[71]. However, label noise is still a challenging task and an open question in computer-aided diagnosis systems. It is arguably made worse by the use of advanced approaches like deep learning, which also typically require a ground truth [69]. The problem is very difficult for mood and psychosis disorder, which share symptoms. It is often challenging to obtain reliable labels for the classification task and in reality, because of many reasons such as insufficient information or even expert’s mistakes or poor data quality we have to deal with labels that are polluted by label noises [38], [39].

Additionally, data acquisition, quality control and preprocessing steps can have an impact on the estimated label noise, though we do careful quality control in order to mitigate these aspects as much as possible. Mislabeled instances may be included in any proportion of data i.e., training, validation and testing dataset which affect the performance of model evaluation and produce unreliable results. Also, there is often considerable overlap among the subjects and no clear boundaries between the groups. For all these reasons and a lack of ground truth in neuroimaging studies, identification of label noise becomes extremely difficult.

Voting filtering is one approach that has been proposed to try to identify label noise and involves removing an instance when all learners agree [38], [71], [72]. We approach this from a data cleansing perspective, that is, we first identify the noisy labels. One of the advantages of a data cleansing approach is it reduces the complexity of the model. It also increases the performance of the classifier; however, as we saw, it may remove a large number of cases. Indeed, if those cases are truly noisy (either because the label or the labeling system is inaccurate or incomplete), then they should be excluded. To avoid over-cleansing problems and mitigating removing minority instances using classification filtering in imbalanced datasets, we used a subsampling approach. We used a random under-sampling approach to handle imbalanced classes and resampled the majority class randomly and uniformly. This was done repeatedly till all the instances of the majority classes were visited at least one time for each cross-validation iteration. Future studies will explore the other possible approaches for addressing imbalanced data.

Simulation study supports our view that one of the consequences of label noise is that it decreases the performance and accuracy of the model. The presence of label noise can disrupt the underlying patterns we are trying to discover. Our analysis of real data also proves this fact that label noise blurred the underlying patterns. The 2D projection using original labels shows lack of similarities between each subject in the 2D space within groups. This does not support the fact that there is a relationship between probands belonging to the same group. However, the cleansed 2D projection shows there is a gradient pattern between groups from healthy controls to the most severe Biotype case B1 or from healthy controls to schizophrenia for DSC. Results were promising, but additional work is needed. For example, some of the results suggest a gradient pattern between groups

from healthy controls to more several patients. Thus the use of additional subcategories or using dimensional information (as suggested in the NIMH RDoC approach) instead of a categorical approach may provide a more meaningful representation of relabeled subjects and underlying relation between them. More detailed mood disorder diagnoses will lead to more effective treatments.

Cleansed data showed many more significant voxels than did the original data. In addition, the DSM-IV cleansed data showed more significant voxels than the Biotype cleansed data. Interestingly, the cleansed dataset showed many more significant differences, providing possible support for our approach. Reclassified relatives had the same distinct features in brain regions more similar to DSM-IV healthy control and Biotype B3 and Biotype normal control in the cleansed dataset.

Reclassifying subjects does not suggest a patient is not actually sick or a healthy person had mental disorders. However, it shows very clearly that categories are not reflecting the underlying biology well. In addition, the fact that 1) the relabeled data showed a clear gradient from the most to least several categories, and 2) there were more voxelwise group differences in regions that were consistent with what might be expected, in the cleansed data, provides intriguing evidence and supports continued work in this direction. Also, the main reason for choosing unsmoothed images was to avoid removing relevant information. Indeed, if we smooth the images, the classification results are lower. Future work can include incorporating multiple types of data (e.g. EEG, structural MRI and functional MRI). In addition, allowing the approach to developing new categories (e.g. via splitting and merging) is another interesting topic for future work. Another interesting avenue of study is to investigate the individuals who were identified as having noisy labels

in more depth (i.e. the boundary cases). Ultimately the results need to be validated clinically.

3.5 Conclusion

In this paper, we proposed a novel approach to estimate label noise prevalent due to incorrect diagnostic classification. We used iterative classification voting filtering using an SVM model. We applied our method to brain imaging in the context of a multi-label imbalanced data from various psychosis disorders and healthy controls. Overall accuracy increases and converges after iterating classification filtering and relabeling steps. The proposed method provides a promising approach for feature extraction of brain images even on noisy datasets by assigning new labels to those inconsistently diagnosed/labeled individuals over multiple iterations. This method identified noisy label samples and suggested new labels for them with the current noisy dataset without requiring extra clean samples in any classification domain by estimating appropriate models and finding optimal hyper-parameters for each classification task. Our method shows that although there was a similar proportion of label noise in multiclass Biotype and DSM-IV categories, label noise distributions were more irregular in DSM-IV than in Biotype categories; however, DSM-IV data showed more significant voxels when evaluating group differences on cleansed data. Our method showed a transparent gradient from the most to least clinically severe groups. Our hope is that this represents an initial step towards a semi-blind categorization approach that is both informed by clinical as well as high-dimensional biological data.

CHAPTER 4. EVALUATION OF BOUNDARIES BETWEEN MOOD AND PSYCHOSIS DISORDER USING DYNAMIC FUNCTIONAL NETWORK CONNECTIVITY (DFNC) VIA DEEP LEARNING CLASSIFICATION

In the upcoming chapter, we delve into the exploration of resting-state fMRI data across various mood and psychosis disorders. Addressing the challenge of valid and reliable diagnoses in psychiatry, especially given the current symptomatic-based categorization, forms the crux of our investigation. In the realm of mental health, there's a growing emphasis on incorporating neurological observations alongside clinical evaluations to enhance diagnostic accuracy. The Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) has pioneered efforts to reclassify psychotic illnesses based on biological markers, deviating from traditional symptomatic measures. However, the correlation between this novel categorization approach and other neuroimaging techniques, particularly resting-state fMRI data, remains relatively unexplored.

Our study aims to bridge this gap by examining the relationship between different psychotic disorders categorization methods and dynamic functional network connectivity (dFNC), a measure derived from resting-state fMRI data. Leveraging cutting-edge artificial intelligence (AI) techniques, we conducted extensive analyses on a dataset comprising 613 subjects, encompassing individuals with psychosis and healthy controls. These subjects were classified using both the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and the B-SNIP biomarker-based (Biotype) approach.

Within each framework, we meticulously conducted statistical analyses, uncovering intriguing disparities in connectivity patterns between individuals classified under DSM-IV and Biotype categories, in comparison to healthy counterparts. Notably, our results revealed distinct transient connectivity states, shedding light on the nuanced differences between various disorders. While both DSM-IV and Biotype categories demonstrated classification accuracy well above chance, the study underscores the importance of further exploration in this direction, including the incorporation of additional data types. These findings present valuable insights, emphasizing the potential benefits of both DSM-IV and biology-based categories while paving the way for future research endeavors in the field.

4.1 Introduction

The validity of psychiatric classification is controversial and remains an open question. The current diagnostic entities in psychiatric nosology for psychosis and mood disorder categorizations have become increasingly untenable and are not biologically based, although presumably, these diagnoses represent categories of brain disorders. The clinical, biological, and overlapping genetic features among psychosis and mood disorder highlight the complexity and lack of clear distinctions among diagnoses within the psychosis dimension [50], [73]–[76]. Studies on three primary entities, schizophrenia, schizoaffective, and psychotic bipolar disorder, demonstrate that individuals meeting the clinical criteria for these diagnoses share many clinical features [74]. The existence of clinical heterogeneity among supposedly distinct categories has led to the suggestion of subtypes within schizoaffective disorder, with one subtype more similar to schizophrenia and the other tending more to resemble bipolar disorder [50]. However, because of poorly specified symptomatic characteristics that can be hard to apply in practice, the reliability

of that subcategorization remains unclear [50]. Because of this unsatisfactory clinical categorization, it has been challenging to identify biological distinctions and biomarkers for DSM-based psychotic disorders.

Multiple ongoing efforts to tackle this problem have attempted to use various biologically-based observations, including neural architecture measures, which might help identify better distinctions among different clinical disorders. The Bipolar-Schizophrenia Network for Intermediate Phenotypes (B-SNIP) [51] is a multisite research consortium focusing on the genetic basis and brain changes associated with psychosis [51]. B-SNIP collects multiple (brain imaging, cognitive, electrophysiologic, oculomotor, genetic) assessment measures within three psychotic disorders: schizophrenia, schizoaffective and psychotic bipolar disorder, and healthy controls. In the B-SNIP study, three new suggested categories ('Biotypes') were introduced based on distinct biological phenotypes obtained from oculomotor, cognitive, and EEG information. These Biotypes cut across traditional diagnostic entities, i.e., each Biotype contains members of all three diagnostic groups. Biotype1 is characterized by subjects with impaired cognitive control and poor sensorimotor function. Biotype2 subjects exhibit impaired cognitive control but exaggerated sensorimotor response. Finally, Biotype3 cases show cognitive and sensorimotor function close to normal [52]. In the most recent B-SNIP study, electroencephalography (EEG) and event-related potentials (ERPs) data analyses were updated, and all B-SNIP subjects were re-quantified based on the updated procedures [77]. The Biotype labels of the B-SNIP subjects which were used in this work are based on these updated categorizations.

Functional magnetic resonance imaging (functional MRI) studies growth in recent years, and various efforts have been conducted on fMRI data to understand human brain mechanisms and diagnosis and treatment of brain disorders [78]. The blood-oxygenation-level-dependent (BOLD) signals measured via fMRI imaging techniques became important tools to quantify and visualize brain activities. fMRI features have been used for studying Schizophrenia, Alzheimer, and many different human brain disorders [46], [48], [79], [80]. Among resting-state fMRI features, functional alteration of the brain is obtained via functional connectivity (FC), widely used to detect abnormalities within the brain and identify biomarkers for brain disorders [1], [78]. Resting-state functional connectivity is the temporal correlation among brain regions obtained using seed-based or data-driven approaches [81]–[84]. Assuming stationary connectivity patterns for the entire scanning period, studies showed that the static functional connectivity could be particularly informative features for the diagnosis and prediction of various disorders, including schizophrenia, and bipolar disorders, from the healthy control group [85]–[87]. More recently, studies tried to address the oversimplified stationary assumption of static functional connectivity and conducted different methods to obtain dynamic functional connectivity patterns [88]–[91]. It has been shown that using the dynamic functional connectivity features can significantly improve the classification performance of mental illness prediction [92], [93].

Analyzing and evaluating DSM-IV and Biotype categories using state-of-the-art artificial intelligence (AI) methods can provide a better understanding of different categorizations and delineate the improvement and shortcomings of new findings. The current nosology has been studied using unsupervised and semi-supervised approaches

applied to structural magnitude resonance imaging (structural MRI) data and showed overlap among different groups [59], [94]. In the current study, we compared the B-SNIP Biotype categories in order to assess brain-based differences compared to the subjects' DSM-IV categories. We used resting-state functional magnetic resonance imaging (fMRI) data (which was not used to define the Biotypes) to evaluate to what extent data from a different biological modality were consistent with the Biotype categories. To do this, we evaluated whether DSM-IV or the Biotype groups aligned better with features extracted from fMRI data. To facilitate comparisons between categories, we performed statistical analysis as well as machine learning techniques on extracted fMRI features based on DSM-IV and biotype labels.

4.2 Materials and Methods

We incorporate multiple steps in our procedure to evaluate time-varying fMRI brain-based differences among the new Biotype and DSM-IV categories. Figure 4-1 shows the schematic diagram of our proposed method, and each part is described in detail in the following sections.

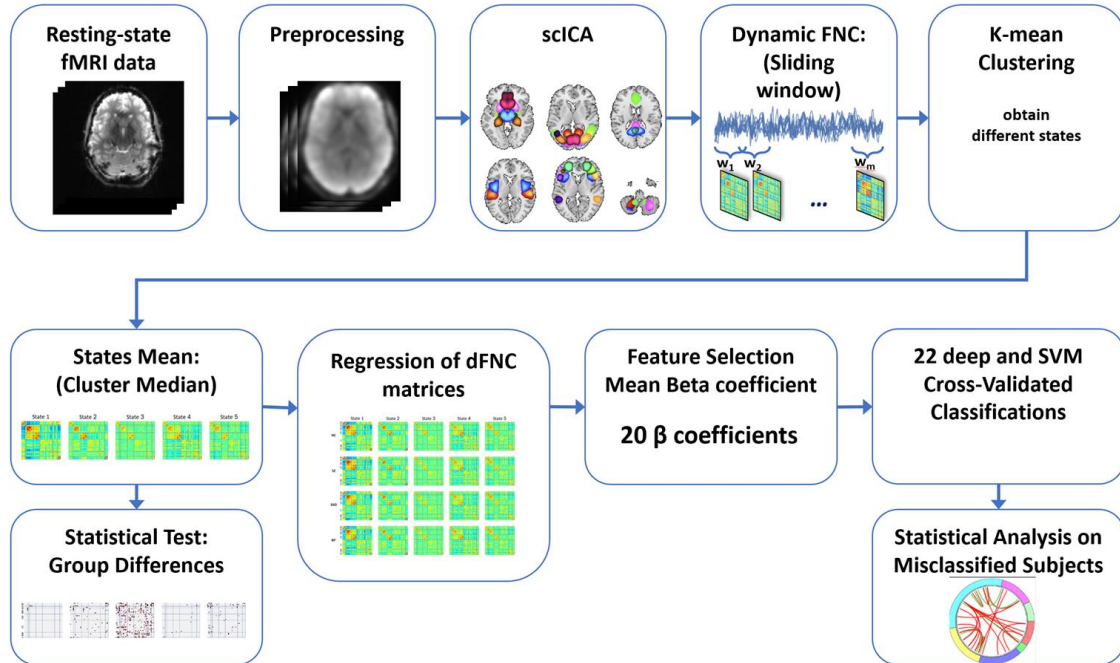


Figure 4-1 – Method diagram of our proposed method. After preprocessing, steps, group independent component analysis using Neuromark pipeline was conducted on resting-state fMRI data and 53 intrinsic connectivity networks (ICNs) extracted, and spatial maps and time courses were estimated for each subject using the back-reconstruction method. Using the window approach, dynamic FNC features were estimated and k-means clustering was applied to dFNC features to estimate FNC states. Statistical analysis and group differences were conducted on dFNC features across diagnoses of DSM biotype categories. From cluster medians of each group and regression analysis, β coefficients were estimated as the features for classification. Comprehensive and various binary and multiclass classification tasks were conducted using the support vector machine (SVM) and the proposed deep models. Statistical analyses were performed between misclassified and correctly classified subjects to extract potential hidden dFNC biomarkers which might add more homogeneity between groups.

4.2.1 Bipolar–Schizophrenia Network on Intermediate Phenotypes Dataset

We analyzed 613 resting-state fMRI datasets collected in B-SNIP-1 (first wave) and re-quantified them using updated electroencephalography (EEG) and event-related potentials (ERPs) analysis procedures. Subjects comprised 131 bipolar, 104 schizoaffective, 155 schizophrenia, and 223 healthy control individuals [77]. All resting-

state fMRI scans were performed on 3T scanners at multiple sites. Additional information about the scanners is described in the supplemental material. The institutional review board at each site approved the study, and all participants provided written informed consent. Subjects were asked to keep their eyes open during the entire scan and to focus on a fixed visual target on display. Participants were diagnosed using two approaches: First, based on the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and all available clinical information (medical charts, information from current treaters) to determine DSM-IV Axis I disorders [51]. Second, participants were classified based on biomarkers obtained from a cognitive battery, stop-signal task, and oculomotor saccadic control task, plus resting EEG and event-related potential data derived from auditory stimulation paradigms [52] into three biological psychosis subtypes, called Biotypes [52]. The demographic information of the subjects is shown in Table 4-1. Each Biotype group significantly contains subjects from all three different DSM-IV categories, as depicted in Figure 4-2.

Table 4-1 – B-SNIP Dataset Demographic

	DSM-IV				Biotype												
	BPP (n=131)		SADP (n=104)		SZP (n=155)		NC (n=223)		B1 (n=129)		B2 (n=121)		B3 (n=140)		NC (n=223)		
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	
Size	131/6 13	21.3 7	104/6 13	16.9 6	155/6 13	25.2 8	223/6 13	36.3 7	129/6 13	21.0 4	121/6 13	19.7 3	140/6 13	22.8 3	223/6 13	36.3 7	
Male	44/13 1	33.5 9	47/10 4	45.1 9	106/1 55	68.3 9	95/22 3	42.6 0	66/12 9	51.1 6	59/12 1	48.7 6	71/14 0	50.7 1	95/22 3	42.6 0	
Female	88/13 1	66.4 1	57/10 4	54.8 1	50/15 5	31.6 1	128/2 23	57.4 0	63/12 9	48.8 4	62/12 1	51.2 4	69/14 0	49.2 9	128/2 23	57.4 0	
Medication																	
No Medication	9/131	6.87	6/104	5.77	9/155	5.80	213/2 23	95.5 1	5/129	3.88	7/121	5.79	12/14 0	8.57	213/2 23	95.5 1	
Psychotropic Medication	120/1 31	91.6 0	98/10 4	94.2 3	145/1 55	93.5 5	9/223	4.03	122/1 29	94.5 7	113/1 21	93.3 9	128/1 40	91.4 3	9/223	4.03	
Unknown Missing	/	2/131	1.53	-	-	1/155	0.65	1/223	0.44	2/129	1.55	1/121	0.82	-	-	1/223	0.44
Sites																	
Chicago	47		21		28		51		28		26		48		51		
Dallas	24		24		20		55		18		26		24		55		
Detroit	11		2		14		20		11		8		8		20		
Hartford	25		33		32		48		27		27		36		48		
Maryland	24		24		61		49		45		34		30		49		
Ethnicity																	
Not Hispanic	119		91		145		203		113		110		132		203		
Hispanic	12		13		10		20		16		11		8		20		
Race																	
African American	27		35		69		65		61		45		25		65		
American Indian	-		-		-		1		-		-		-		1		
Asian	4		1		3		11		3		1		4		11		
Caucasian	95		62		77		138		59		72		103		138		
Native Hawaiian	-		-		-		1		-		-		-		1		
Multiracial Mixed Race	/	1	5		5		3		4		2		5		3		
Other Race	4		1		1		4		2		1		3		4		
Unknown Missing	/	-	-		-		-		-		-		-		-		
	<i>Mea n</i>	<i>Std</i>	<i>Mea n</i>	<i>Std</i>	<i>Mea n</i>	<i>Std</i>	<i>Mea n</i>	<i>Std</i>	<i>Mea n</i>	<i>Std</i>	<i>Mea n</i>	<i>Std</i>	<i>Mea n</i>	<i>Std</i>	<i>Mea n</i>	<i>Std</i>	
Age	36.13	±12.56	35.65	±12.07	35.77	±12.32	38.23	±12.58	36.48	±12.46	37.57	±11.97	33.80	±12.25	38.23	±12.58	

Table 4-2 – continued

PANSS																
<i>Positive</i>	12.54	±4.26	17.94	±5.22	16.73	±5.62	-	-	15.22	±5.44	17.46	±5.64	14.52	±5.25	-	-
<i>Negative</i>	12.10	±3.73	15.52	±4.76	16.63	±5.98	-	-	14.65	±5.14	15.31	±5.44	13.53	±5.48	-	-
<i>General</i>	28.42	±7.95	34.45	±9.08	31.82	±8.86	-	-	31.40	±9.13	32.63	±8.31	31.23	±9.16	-	-
<i>Total</i>	53.06	±13.27	67.85	±16.43	65.19	±17.17	-	-	60.20	±16.99	65.41	±16.42	60.28	±16.99	-	-



Figure 4-2 – Number of subjects for each group of DSM-IV and Biotype categories of B-SNIP dataset. The percentage in front of each group shows how each proband group diagnosed as other categories.

4.2.2 Preprocessing and ICNs extraction

fMRI data were preprocessed using statistical parametric mapping (SPM 12 <http://www.fil.ion.ucl.ac.uk/spm/>). To correct subjects' head motion, a rigid body motion correction was performed, followed by the slice-timing correction to account for the timing

difference in slice acquisition. The fMRI data were subsequently warped into the standard Montreal Neurological Institute (MNI) space using an echo planar imaging (EPI) template. Data were resampled to $3 \times 3 \times 3$ mm³ isotropic voxels and smoothed using a Gaussian kernel with a full width at half maximum (FWHM) of 6mm. After preprocessing steps, 53 intrinsic connectivity networks (ICNs) were extracted using a fully automated Neuromark ICA pipeline [95], which extracts comparable features (ICNs) for each subject based on prior knowledge. The Neuromark priors were created by estimating a brain network template using group independent component analysis (ICA) applied separately to resting-state functional magnetic resonance imaging (fMRI) data from two independent datasets containing 1828 controls and identifying the replicated intrinsic networks. These networks were ordered and labeled and used as spatial constraints (priors) in a single-subject spatially constrained ICA run on each of the B-SNIP datasets. The spatially constrained ICNs make it more feasible to compare subjects from different studies, sites, and disorders. Finally, we performed quality control (QC) as described in [95] to ensure high-quality data.

4.2.3 dFNCs and Exemplar clustering

For each subject, dynamic functional connectivity (dFNC) was calculated using a tapered sliding windows approach [88]. The inverse of covariance matrices is estimated using graphical lasso in the cross-validated form to optimize the regularization parameter lambda for each window [96]. Then, Fisher transformation was applied to dFNCs data, and the k-means clustering algorithm was used on windowed covariance matrices to obtain dFNC patterns (states) that occurred across the time and subjects. First, those windows which had FC variance local maxima were chosen as subject exemplars. Then, a k-means clustering algorithm with random initialization was performed on subject exemplars to

obtain centroid as initialization of the next step, clustering of all windowed FNCs. The optimal number of clusters was obtained using the elbow criterion of the cluster validity index, which is computed as the ratio of within-cluster distances to between-cluster distances. The optimal number was determined as 5, resulting in 5 dFNC states, which is consistent with previous dFNC studies [88], [92], [97]–[99].

4.2.4 *Group differences and State Occupancies*

To assess differences between groups, we performed a cell-wise group-level analysis of variance (ANOVA) on functional connectivities for each state. FDR corrections were performed on the statistical test results, and significant pairs of components were identified at the significant level $\alpha < 0.05$ for both categories DSM-IV and Biotypes for each state and visualized using obtained F-values. In addition, we analyzed how different groups from both categories were placed into different dFNC states.

4.2.5 *Classification*

We performed classification on various binary and multi-class tasks using both our proposed deep learning model and support vector machine (SVM) classifier. This classification comprised six binary classification tasks, four 3-way classification tasks, and one 4-way classification task, totaling 22 total classification tasks for two categories, DSM-IV and Biotypes. Table 4-3 shows details of the classification tasks. We first extracted the features for classification models, such that for each group, 5 cluster medians were obtained, one for each state, and concatenated these into a tensor with dimensions $4 \times 5 \times 53 \times 53$. Next, using the ordinary least squares (OLS) linear regression model, we regressed these against each windowed FNC matrices for each subject to obtain

standardized beta coefficients for each state. These beta coefficients were then averaged for each subject, totaling 20 such measures per individual. This approach was previously introduced for a three-way classification problem elsewhere [92]. Next, these features were used for classification using the deep and SVM models.

Table 4-3 – Classification Tasks

c	Classification Tasks	
	DSM-IV	Biotype
2-way	NC Vs. BP	B1 Vs. NC
	NC Vs. SAD	B2 Vs. NC
	NC Vs. SZ	B3 Vs. NC
	SZ Vs. SAD	B1 Vs. B2
	SZ Vs. BP	B1 Vs. B3
	SAD Vs. BP	B2 Vs. B3
3-way	NC Vs. SZ Vs. SAD	B1 Vs. B2 Vs. NC
	NC Vs. SAD Vs. BP	B1 Vs. B3 Vs. NC
	NC Vs. SZ Vs. BP	2B Vs. 3B Vs. NC
	SZ Vs. SAD Vs. BP	B1 Vs. B2 Vs. B3
4-way	NC Vs. SZ Vs. SAD Vs. BP	B1 Vs. B2 Vs. B3 Vs. NC

4.2.6 Deep model Classification

The advancement of deep neural networks in various applications indicates that it is a powerful machine learning framework. Deep neural networks can represent complex functions, and their complexity can increase by having more layers and units [100], [101].

Recent studies show the potential of deep learning models in brain imaging data analysis [1], [32]. The capability of deep learning to detect and learn complex and nonlinear patterns made it an attractive tool in the field of neuroimaging studies and mental illness. In this study, we performed a multi-layer feed-forward network model that combined the FNC beta features with the sex, age group, and categorical site features as input to our deep model, which consisted of 4 consecutive fully connected layers with sizes of 256, 128, 64, and 32 units, each equipped with a ReLU (Rectified Linear Unit) activation function and 0.5 dropout layer between them to learn the non-linearity of the input features and to avoid over-fitting. ReLU units are easy to optimize and more useful for learning process and because of their derivative property they are usually an acceptable choice for hidden units activation function [100]. Stacking multiple layers of fully connected layers and increasing the number of parameters in deep models results in better performance [100]. In addition, it has been shown that the dropout technique reduces the problem of overfitting and improves generalization performance [102]. Lastly, we used a fully connected layer with the size of the number of classes and a SoftMax layer to obtain the probability for each class for each classification task. The architecture of our deep model is shown in Figure 4-3. For training the model, we utilized the Adam algorithm [103] for stochastic gradient optimization and the early stopping method to cut off early and prevent over-fitting.

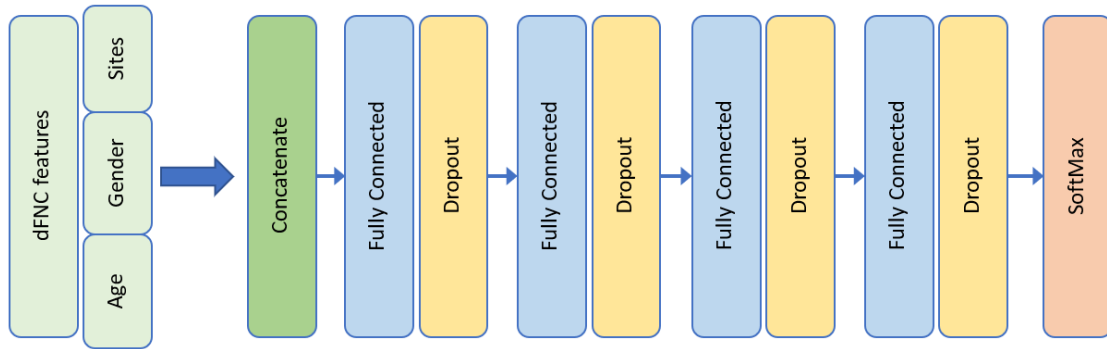


Figure 4-3 – Deep classification model architecture.

4.2.7 SVM Classification

In addition to the deep model, we also performed classification using the dynamic FNC matrices [92] via an SVM approach. To mitigate against bias, we cross-validated the results. Grid search hyperparameter tuning was performed for the SVM model to obtain the best hyperparameter. We used a linear and radial basis function (RBF) kernel for our grid search domain with the range of (0.0001, 10000.0) and (0.0001, 1) for the regularization parameter and kernel coefficient for the RBF kernel, respectively.

4.2.8 Classification setup

We performed a nested cross-validation setup in order to optimize the hyperparameters and mitigate the risk of the data leakage. It has been shown that nested cross-validation reduces the potential over-fitting and improves the generalization performance [104]. As outer cross-validation and to optimize generalization performance, the data for each task was segmented into training and test sets using 5-fold stratified cross-validation. Then for each iteration, the remained k-1 fold train data was also segmented

using inner 3-fold stratified cross-validation into train and validation set to optimize parameters and hyperparameters. After selecting the best model in the inner cross-validation, the performance of the model was evaluated as the average of the accuracies across testing sets.

4.2.9 *Classification analysis*

To obtain more detailed information about the dataset and to better understand the complexity of each category, we performed an analysis of the whole dataset using the probability on the test folds obtained from the classification model. First, we performed group difference statistical tests of states' dFNC matrices from correctly predicted subjects and misclassified ones. For each pair of groups, two statistical tests were performed using misclassified subjects and the other correctly classified subjects. One test was performed to find the differences between misclassified subjects and their primary (given labels) group, and the second to find the difference between misclassified subjects and their predicted group, which were classified correctly by the model. The purpose of these comparisons was to see how misclassified subjects differed from the two groups of primary (given labels) and suggested (predicted labels). We performed cell-wise statistical t-tests to obtain dFNC significant differences between groups for each pair for each state. For 1378 pairs of components, FDR correction was applied for each state using a significance level of 0.05.

We performed additional analysis on misclassified subjects using the probability of the models which were assigned to each group to visualize the result of the model into a 3-dimensional barycentric coordinate system for both DSM-IV and Biotype groupings. The

barycentric scatter plot (Figure 7) shows the relation of four classes using the probabilities assigned to subjects from the model with higher performance. Each subject could be considered as a point with four probability values per class, with points then transferred into the unique convex coordination of the four vertices and visualized in a barycentric coordinate system. The dots close to the vertices show a high probability of that class; consequently, dots close to the middle (barycenter) show similar probability values for all classes. This representation better conveys the underlying pattern obtained via the classifier of the dataset.

4.3 Results

4.3.1 Intrinsic connectivity networks and dynamic functional connectivity

The 53 intrinsic connectivity networks (ICNs) estimated using the Neuromark pipeline [31] were grouped into seven network domains, including sub-cortical (SC), auditory (AU), visual (VIS), sensorimotor (SM), cognitive control (CC), default mode (DM) and cerebellar (CB). Dynamic functional connectivity estimated via a windowed approach [88] using the time-courses of these 53 ICNs produced a 3-dimensional, $613 \times 178 \times 1378$ tensor. Five dynamic functional connectivity states were estimated using the exemplar method [88] for each group. The group average functional network connectivity states are shown in Figure 4-4.

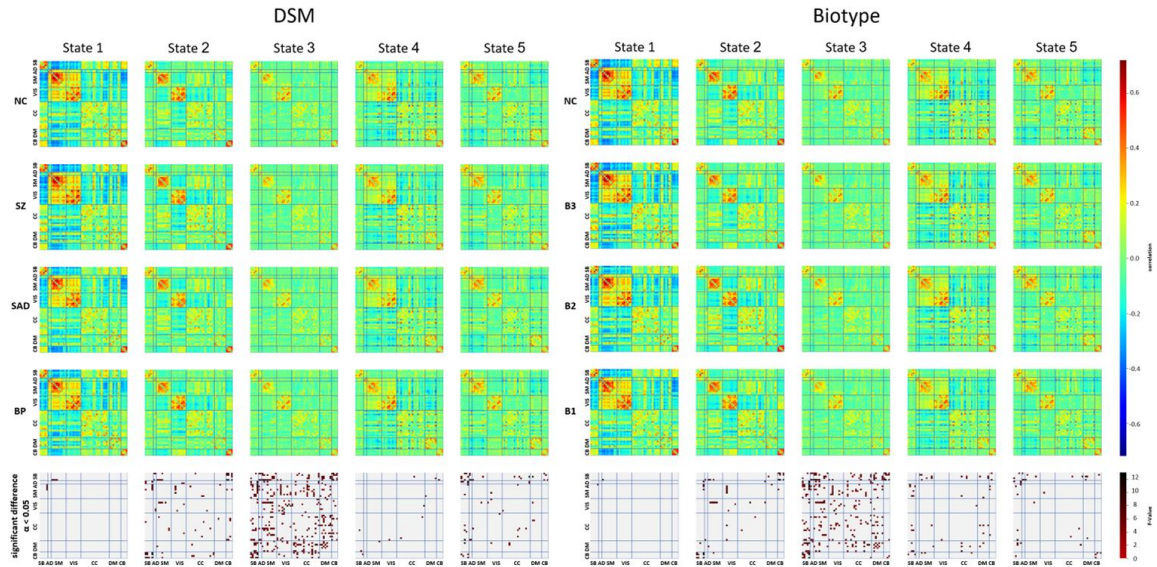


Figure 4-4 – Dynamic states of 4 groups of DSM-IV and Biotype categories averaged across each state’s windowed dNFC. The result of the ANOVA statistical test and identified significant pair of connections are highlighted in the last row.

Some main differences in connectivity patterns are observed between dFNC states. State 1 and State 2 show moderate to high correlation within SM and VIS domains’ ICNs. In state 1, SB regions are negatively moderate to highly correlated with VIS, SM, and AD, which cannot be observed in the other four states. State 1 shows the strongest connectivity pattern both within and between VIS, SM, and AD. State 3 is the weakest among other states. State 4 distinguishes itself from state 3 with respect to the connectivity of DM ICNs; connectivity within DM regions is stronger in state 4. In addition, more anticorrelated from VS, SM, and AD regions can be seen in state 4. State 5 shows more anticorrelated between SM and AU regions and CB regions in comparison with state 4, and the moderate correlation between VIS and SM that can be seen in state 4, is not the case in state 5.

Table 4-4 shows how each group occupied different states. Besides state3 occupancy, which was the most occupied state among all groups, the occupancy of healthy controls in

the state1 was significantly higher in comparison with all patient groups of both Biotypes and DSM-IV. In state2, which was the least occupied state for controls, the occupancy among patient groups was significantly higher in comparison with controls. The pattern of occupancies among patients was more consistent with the Biotype classification. For example, in the state1, occupancies of Biotype groups decreased from 8% in Biotype3 to 6% in Biotype2 and Biotype1. In state2, Biotype occupancies were more regular in accordance with the severity of impaired cognitive functions observed in the patients (i.e., most abnormal in Biotype1, least abnormal in Biotype3). Controls and Biotype3 subjects shared fewer patterns in state2 in comparison with Biotype1 and Biotype2. The result also shows the schizophrenia group had more patterns in state2. In general, the occupancies of Biotype3 are more similar to controls than to either Biotype1 or Biotype2.

Table 4-4 – States’ Occupancy

Type	State1	State2	State3	State4	State5	Occupancy State order
DSM-IV						
SZP	6%	21%	38%	19%	17%	3 > 2 > 4 > 5 > 1
SADP	8%	16%	43%	16%	16%	3 > 2 > 4 > 5 > 1
BPP	6%	14%	42%	20%	18%	3 > 4 > 5 > 2 > 1
NC	13%	10%	40%	21%	16%	3 > 4 > 5 > 1 > 2
Biotype						
B1	6%	18%	41%	19%	16%	3 > 4 > 2 > 5 > 1
B2	6%	21%	40%	17%	16%	3 > 2 > 4 > 5 > 1
B3	8%	12%	42%	20%	19%	3 > 4 > 5 > 2 > 1
NC	13%	10%	40%	21%	16%	3 > 4 > 5 > 1 > 2

Statistical group differences for each state indicate that Biotype groups have fewer significant component pairs in comparison with DSM-IV groups. The result of the ANOVA statistical test and identified significant pair of connections after FDR correction at the significant level $\alpha < 0.05$ are depicted in Figure 4-4 (last row). The number of significant pairs for each state is also shown in Table 4-5.

Table 4-5 – Number of significant pair obtained from ANOVA statistical test

Type	State1	State2	State3	State4	State5	Total
DSM	4	41	174	6	29	254
Biotype	1	25	134	25	18	203

Functional connectivity differences between groups were observed in all states. The results in state2, as the second most occupied state by schizophrenia and schizoaffective, show greater connectivity in healthy controls than in schizophrenia and schizoaffective groups between the cerebellum and subcortical components, including the subthalamus and putamen. In addition, healthy groups showed greater connectivity than all patients between the cerebellum, the caudate, and the thalamus components. Larger hyperactivation in normal controls was observed between the thalamus and sensorimotor regions, including the postcentral gyrus, right postcentral gyrus, left postcentral gyrus, and paracentral lobule. Similarly, normal control showed greater hyperactivation between the postcentral gyrus (SM) and posterior cingulate cortex (DM), right postcentral gyrus (SM) and cerebellum (CB), and hippocampus (CC) and cerebellum (CB). This implies that schizophrenia and

schizoaffective groups tend to transition in the state2 more than other groups, which there is impaired FC between the cerebellum and different brain regions, including SB (subthalamus, putamen, caudate, thalamus). Prior studies showed similar findings regarding disconnection between the cerebellum and other brain regions in schizophrenia patients [91], [105], [106].

4.3.2 *Classification results*

Classification accuracy is depicted in Figure 4-5, and performance metrics, including average accuracy, precision, recall, and area under the curve (AUC) of 22 classification tasks, are shown in Table 4-6. In addition, confusion matrices of 3-way and 4-way classification tasks were obtained from cross-validation runs shown in Figure 4-6. Binary classification tasks were the most accurate models compared to multi-class, 3-way, and 4-way classification tasks. Scores for binary classification tasks ranged from 0.84 to 0.72. The most distinctive classification was obtained using the deep model from DSM-IV categories between controls and schizoaffective patients (NC Vs. SAD) with an accuracy score of 0.84. The lowest accuracy of 0.72 was the Biotype comparison between Biotype2 and Biotype3 proband. The 3-way classification task scores were from 0.60, which were obtained from the three Biotype categories (B1 Vs. B2 Vs. B3) task to 0.72 for the normal, schizoaffective, and bipolar group (NC Vs. SAD Vs. BPP) task using the deep model. The situation was similar for 4-way classification tasks, with a 0.60 score for DSM-IV and 0.58 for Biotypes.

Table 4-6 – Deep model classification results

Classification Tasks										
k-way	DSM-IV					Biotype				
	Task	Precision	Recall	ACC	AUC	Task	Precision	Recall	ACC	AUC
2-way	NC Vs. BP	0.78	0.78	0.78	0.77	B1 Vs. NC	0.83	0.83	0.83	0.9
	NC Vs. SAD	0.83	0.83	0.83	0.79	B2 Vs. NC	0.84	0.83	0.83	0.92
	NC Vs. SZ	0.83	0.83	0.83	0.82	B3 Vs. NC	0.77	0.77	0.77	0.84
	SZ Vs. SAD	0.79	0.79	0.79	0.78	B1 Vs. B2	0.76	0.76	0.76	0.76
	SZ Vs. BP	0.75	0.76	0.76	0.75	B1 Vs. B3	0.72	0.72	0.72	0.78
	SAD Vs. BP	0.77	0.77	0.77	0.76	B2 Vs. B3	0.72	0.72	0.72	0.72
3-way	NC Vs. SZ Vs. SAD	0.73	0.73	0.73	0.85	B1 Vs. B2 Vs. NC	0.7	0.7	0.7	0.85
	NC Vs. SAD Vs. BP	0.7	0.71	0.71	0.83	B1 Vs. B3 Vs. NC	0.66	0.67	0.67	0.82
	NC Vs. SZ Vs. BP	0.66	0.67	0.67	0.85	B2 Vs. B3 Vs. NC	0.64	0.64	0.64	0.82
	SZ Vs. SAD Vs. BP	0.64	0.64	0.64	0.81	B1 Vs. B2 Vs. B3	0.6	0.59	0.59	0.76
4-way	NC Vs. SZ Vs. SAD Vs. BP	0.6	0.6	0.6	0.81	B1 Vs. B2 Vs. B3 Vs. NC	0.57	0.58	0.58	0.78

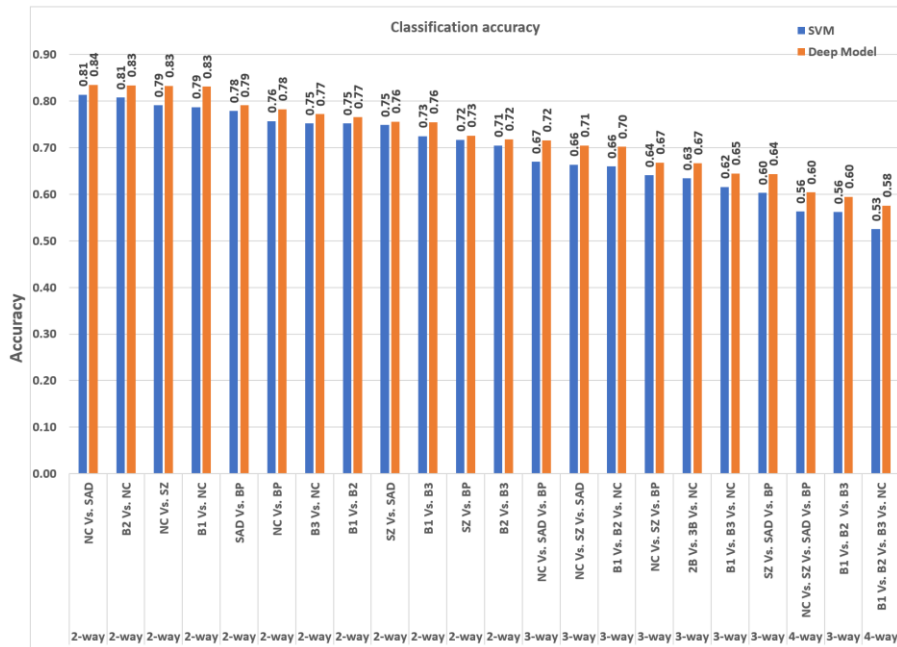


Figure 4-5 – Classification accuracy of the total 22 tasks, including using deep and SVM models. For each task, the average accuracies obtained from cross-validation

runs of SVM and deep models are shown. The tasks sorted from the highest accuracy (left side) belong to the DSM-IV binary task between the normal control group and schizoaffective group to the lowest accuracy (right side) of the biotype 4-way classification task.

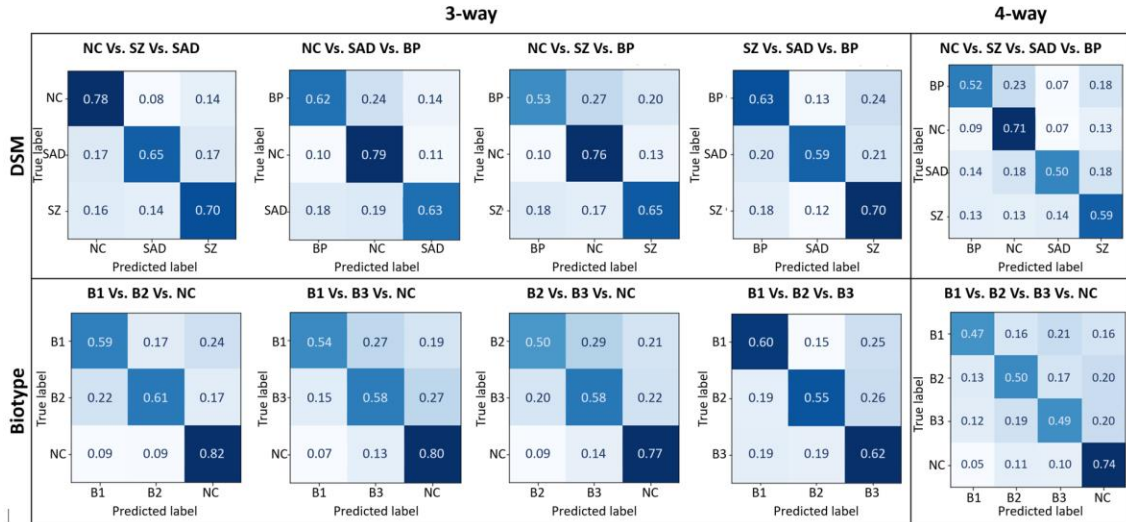


Figure 4-6 – Confusion matrices of 3-way and 4-way classification tasks of DSM-IV and biotype categories obtained from deep model. The confusion matrices were normalized over the true (given) labels and by rows.

4.3.3 Classification analysis

Testing over state dFNC matrices in 4-way classification tasks revealed that between misclassified subjects and their original given labels in the DSM-IV category, a total of 91 ICN pairs showed significant differences. The number of total differences between misclassified DSM-IV subjects and their congruent suggested group, which were classified correctly, was 23. For the Biotype categories, 161 ICN significant difference pairs were identified between misclassified subjects and their primary corrected classified group, and 36 ICN significant difference pairs were obtained between misclassified subjects and their suggested group. Among the five different states, state 1 had the highest number of differences between the model predicted label group and misclassified subjects. The model

confusion was highest in state 1 between Biotype1 and controls. Statistical tests of state dFNC matrices of misclassified subjects are depicted in Figure 4-7. The total number of the differences are summarized in Table 4-7, and the numbers in group levels in Table 4-8.

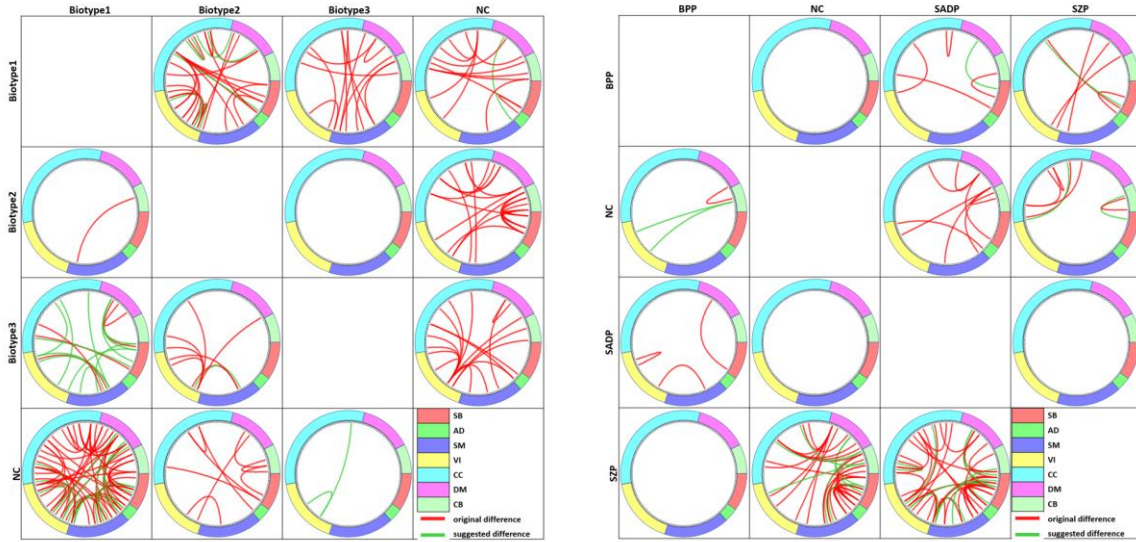


Figure 4-7 – The connectogram (chord diagram) of group difference between misclassified subjects and their original and suggested group by the model. Rows are the original (given labels and the columns are the suggested labels. Each cell belongs to a pair of given label group and predicted label group based on the relation of their differences with the misclassified subjects. Red links shows significant different FC pair between misclassified subjects and their original groups and green links significant different FC pair between misclassified subjects and their suggested correctly classified groups.

Table 4-7 – Total number of significantly different FC pairs of misclassified subjects with their original and predicted group

	DSM						Biotype					
	State1	State2	State3	State4	State5	All States	State1	State2	State3	State4	State5	All States
Original	59	15	6	8	3	91	83	9	33	20	16	161
Suggested By Model	16	7	-	-	-	23	32	1	-	3	-	36

Table 4-8 – Number of significant FC pairs of misclassified subjects with their original and predicted group

DSM									Biotype								
Group pair (given Vs. predicted as)			Number of significant FC pairs						Group pair (given Vs. predicted as)			Number of significant FC pairs					
			sta te1	sta te2	sta te3	sta te4	sta te5	all states				sta te1	sta te2	sta te3	sta te4	sta te5	all states
BPP Vs. SAD P	given Vs. predicted	misclassified	4	-	-	-	-	4	Biotype pe1	given Vs. predicted	misclassified	23	-	-	-	-	23
SAD P Vs. BPP	predicted Vs. misclassified		0	1	-	-	-	1	Biotype pe2	given Vs. predicted	misclassified	6	-	-	-	-	6
BPP Vs. SZP	given Vs. predicted	misclassified	1	-	-	1	3	5	Biotype pe1	given Vs. predicted	misclassified	6	-	1	1	5	13
SZP Vs. BPP	predicted Vs. misclassified		1	-	-	-	-	1	Biotype pe3	given Vs. predicted	misclassified	-	-	-	-	-	0
BPP Vs. NC	given Vs. predicted	misclassified	-	-	-	-	-	-	Biotype pe1	given Vs. predicted	misclassified	-	-	8	4	-	12
NC Vs. SAD P	predicted Vs. misclassified		-	-	-	-	-	-	Biotype pe2	given Vs. predicted	misclassified	1	-	-	-	-	1
SAD P Vs. BPP	given Vs. predicted	misclassified	-	3	-	-	-	3	Biotype pe1	given Vs. predicted	misclassified	-	-	1	-	-	1
BPP Vs. SAD P	predicted Vs. misclassified		-	-	-	-	-	-	Biotype pe2	given Vs. predicted	misclassified	-	-	-	-	-	-
SAD P Vs. SZP	given Vs. predicted	misclassified	-	-	-	-	-	-	Biotype pe3	given Vs. predicted	misclassified	-	-	-	-	-	-
SZP Vs. NC	predicted Vs. misclassified		-	-	-	-	-	-	Biotype pe1	given Vs. predicted	misclassified	-	-	-	-	-	-
SAD P Vs. NC	given Vs. predicted	misclassified	-	-	-	-	-	-	Biotype pe2	given Vs. predicted	misclassified	-	-	21	-	-	21
NC Vs. SZP	predicted Vs. misclassified		-	-	-	-	-	-	Biotype pe3	given Vs. predicted	misclassified	-	-	-	-	-	-
SZP Vs. BPP	given Vs. predicted	misclassified	-	-	-	-	-	-	Biotype pe1	given Vs. predicted	misclassified	3	-	1	-	1	5
BPP Vs. SZP	predicted Vs. misclassified		-	-	-	-	-	-	Biotype pe2	given Vs. predicted	misclassified	12	-	-	-	-	12
SZP Vs. SAD P	given Vs. predicted	misclassified	33	2	-	-	-	35	Biotype pe3	given Vs. predicted	misclassified	-	2	-	5	-	7
SAD P Vs. SZP	predicted Vs. misclassified		10	-	-	-	-	10	Biotype pe1	given Vs. predicted	misclassified	-	1	-	-	-	1
SZP Vs. NC	given Vs. predicted	misclassified	13	5	5	7	-	30	Biotype pe2	given Vs. predicted	misclassified	2	-	-	10	8	20
NC Vs. BPP	predicted Vs. misclassified		3	4	-	-	-	7	Biotype pe3	given Vs. predicted	misclassified	-	-	-	-	-	-
NC Vs. BPP	given Vs. predicted	misclassified	-	-	1	-	-	1	NC Vs. Biotype pe1	given Vs. predicted	misclassified	49	-	1	-	-	50
BPP Vs. NC	predicted Vs. misclassified		2	-	-	-	-	2	Biotype pe2	given Vs. predicted	misclassified	13	-	-	1	-	14
NC Vs. SAD P	given Vs. predicted	misclassified	8	-	-	-	-	8	Biotype pe3	given Vs. predicted	misclassified	-	7	-	-	2	9
SAD P Vs. NC	predicted Vs. misclassified		-	-	-	-	-	-	Biotype pe1	given Vs. predicted	misclassified	-	-	-	-	-	-
NC Vs. SZP	given Vs. predicted	misclassified	-	5	-	-	-	5	NC Vs. Biotype pe2	given Vs. predicted	misclassified	-	-	-	-	-	-
SZP Vs. NC	predicted Vs. misclassified		-	2	-	-	-	2	Biotype pe3	given Vs. predicted	misclassified	-	-	-	2	-	2

The barycentric scatter plot shows the probability of each subject being assigned to a tetrahedron on the 3D surface. (3D barycentric coordinates were obtained from the probability of each subject estimated from the deep model, which provided better

performance). Figure 4-8 depicts the correctly classified (left) and misclassified (right) subjects for DSM-IV and Biotype analyses. The higher probability the subjects obtain from a class, the more dots are close to that vertex. The Figure shows how the misclassified subjects are placed in different groups.

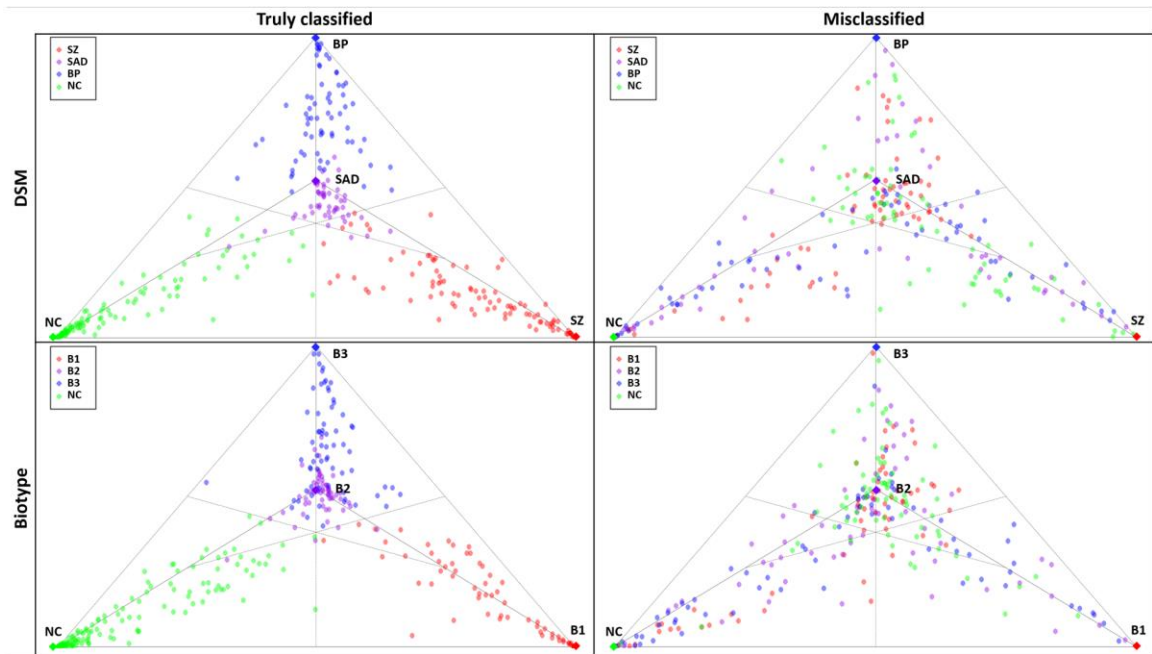


Figure 4-8 – 3D tetrahedron scatter plot of the probability values of correctly classified (left) and misclassified subjects (right) for DSM group(top) and Biotype group (bottom). The dots close to the vertices have a higher probability for that class from the model.

4.4 Discussion

In this study, we evaluated dynamic functional network connectivity measures across psychosis participants in addition to healthy controls, categorized using conventional DSM-IV diagnosis and an alternative B-SNIP-derived Biotype categorization. We provided statistical analysis and group differences of dynamic states and analysis of state occupancies as a complementary investigation to classification analysis. This

complimentary analysis enables us to compare our findings with previous studies in the literature that investigated similar data standard statistical approaches. Moreover, using classification results, we conducted post hoc statistical analysis to investigate heterogeneity within and across psychotic disorders based on functional MRI dynamic states. Using the combination of classifications and statistical analyses, we identified and revealed data-driven contributing features for the misclassified subjects in different tasks which might be informative for a better understanding of the heterogeneity across mental disorders. Previous studies mostly focused on statistical analyses to study the differences across mood and psychosis groups.

Among binary classification tasks, some of the Biotype groups showed better performance in comparison with DSM-IV; however, there were some cases (e.g., Biotype2 Vs. Biotype3 and Biotype1 Vs. Biotype3) that had a lower performance with respect to some of the DSM-IV.

Statistical differences of misclassified subjects with their primary and adjusted categories showed that functional connectivity patterns of the model-predicted group were more homogeneous in comparison with the original DSM-IV and Biotype categories. The tetrahedron plots show that misclassified subjects do not follow any clear pattern, which supports the idea that there is a lack of clear boundaries between different groups in current nosology. This also suggests that classification based on a continuous spectrum could be more meaningful. In most cases, the differences between misclassified subjects and their original groups were higher than the differences between them and their adjusted groups. However, some scenarios show heterogeneity still exists between the misclassified subjects and their adjusted groups.

In 4-way classification tasks that include all groups, misclassified bipolar patients revealed functional connectivity patterns close to schizophrenia in different brain regions including, the sensorimotor and default mode domains, the visual and default mode, and the sub-cortical and the auditory domains. The other misclassified bipolar subjects who were predicted as schizoaffective revealed differences with their original groups in their functional connectivity measures between the sub-cortical and the cognitive control domains, the sub-cortical and the cerebellar domains, and cognitive control and visual domains. Schizophrenia patients that were predicted as controls had more similar patterns close to controls and revealed many differences between their original schizophrenia groups and almost between all domain regions. The situation for misclassified schizophrenia patients who were predicted as schizoaffective also showed many differences between all domain regions. The normal control subjects who were predicted as schizoaffective revealed functional connectivity differences with other normal control subjects between the default mode network and the cognitive control, the default mode network and the sub-cortical, the default mode network and visual domains, as well as cerebellar and sensorimotor and visual domains. In addition, healthy control who were classified as schizophrenia revealed differences in their functional connectivity with their own group among the cerebellar and the subcortical domains as well as the cognitive control regions. These differences could not be detected when we compared those with the correctly classified schizophrenia group which indicates that their patterns were more similar.

The result of the classification task between healthy controls and Biotype1 showed the most significant difference between misclassified subjects which were classified as

Biotype1, and their primary group, normal control. This same issue also occurs between Biotype2 misclassified subjects and model-suggested group Biotype1, suggesting that in some cases, the subjects labeled as Biotype2 and Biotype1 did not clearly follow the pattern of either group on the basis of fMRI dFNC measures. Thus, the analysis shows that overlap still exists between different groups. The problem is not addressed by categorical approaches such as DSM-IV and Biotypes, and the differences between groups are not clearly defined, which reveals a challenging problem in classification and nosology. Introducing new subtypes considering additional biological findings not used in defining the original B-SNIP Biotypes could be one possible solution to this issue.

In the B-SNIP study, although the identifying biomarker measures were gathered on the control group and used to establish which biological data separate patients from controls better, they were not used subsequently to classify the healthy controls into Biotypes. Misclassification of different patient subjects of both DSM-IV and Biotype categories as normal controls and misclassification of some normal subjects as belonging to different patient groups (e.g., normal controls misclassified as Biotype1 or as schizophrenia) suggests that the study of all groups, including controls, might improve the homogeneity of these categories. Not including the healthy control in three Biotype groups might negatively impact this categorization and decrease the homogeneity of this category. Our results suggest that considering healthy individuals in Biotype nosology might be more helpful, and conceivably some of these might be grouped into different categories similar to the patients.

The overall performance of Biotype 3-way classification tasks, which includes all three Biotype groups, shows the lowest accuracy among all other 3-way classification

tasks, and the score is lower than the 4-way classification task in DSM-IV, suggesting that the new Biotype categorization of might need updating by including other modalities and introducing subtypes which involves healthy controls.

One of the limitations of this study is that we used only feature extraction for analyzing the fMRI data. The lack of access to ground truth makes psychosis and mood disorder nosology a challenging and complex question. Incorporating additional neuroimaging data modalities such as genomic, structural MRI data and even other resting-state fMRI features might improve our understanding of the shortcomings of the current categorizations and also allow for the optimization of biomarkers for nosology and defining more homogeneous subtypes. Previous studies using unimodal structural MRI data and grey matter analysis for affective and non-affective psychoses, including schizophrenia, bipolar disorder, and schizoaffective categories showed revealed accuracy is between 59.50% and 73.00% [9]–[12] Insufficient sample size in different studies caused inconsistencies in finding across bipolar and schizophrenia groups, and the results require further validation, [1], [18]. The vast majority of the research focused on binary classification studies of single disorder and control. Indeed, multi-class classification tasks across psychosis and control groups got less attention. One of the big challenges in psychosis categorization is the overlapping symptoms across groups and subgroup diagnosis, which we will investigate in more detail in future work. But it is worth mentioning that single case versus control studies ignore the overlapping problem in psychiatry and the heterogeneity across groups [1].

In addition, further investigation should be conducted via different machine learning and deep learning methods. Tree ensemble machine learning approaches such as gradient

boosted decision trees (GBDT) and XGBoost are widely used on tabular data and might improve performance. In addition to standard machine learning approaches, several studies attempted to investigate deep learning models in tabular data [107].

4.5 Conclusion

We analyzed mood and psychosis disorder classification with state-of-the-art methods in neuroimaging and AI on a relatively large dataset and compared standard symptom-based DSM-IV categorization with a biologically-based Biotype categorization approach. We obtained dynamic functional network connectivity from resting-state fMRI data and analyzed the FNC patterns using SVM classification applied to the extracted features. Results showed evidence in support of both categorization approaches, as well as limitations. Importantly, there is a challenge with overlap among different groups of patients and controls in a multi-class problem. The general approach of using biologically based measures such as neuroimaging to inform categories is a promising direction for future nosology of mood and psychosis disorders, but more work remains to be done.

CHAPTER 5. A DEEP LEARNING APPROACH FOR PSYCHOSIS SPECTRUM LABEL NOISE DETECTION FROM MULTIMODAL NEUROIMAGING DATA

In this chapter, we embark on a comprehensive exploration of the intricate structural and functional mechanisms underlying mood and mental disorders, employing advanced neuroimaging techniques. The limitations inherent in individual modalities have long constrained our understanding of these disorders. Compounding this challenge is the lack of validity in current psychosis nosology, further complicating the identification of biomarkers.

Our study introduces innovative deep convolutional frameworks designed to classify and detect label noises within structural and functional magnetic resonance imaging (MRI) data. Leveraging these frameworks, we conducted an in-depth analysis utilizing functional and structural MRI data sourced from a Schizophrenia dataset. Through meticulous evaluation in a cross-validated manner, we assessed the model's performance. Additionally, we introduced a novel noise criterion, enabling the identification of potentially noisy subjects for each modality.

The outcomes of our study revealed compelling insights. Specifically, we found that the model trained on resting-state functional MRI data outperformed its structural MRI counterpart, demonstrating superior informativeness. Furthermore, by scrutinizing the noise levels, we identified potential borderline subjects, hinting at potential subtypes within

the dataset. Employing rigorous statistical analysis, we discerned differences in resting-state static functional connectivity features, shedding light on these intricate subtypes.

This chapter offers a deep dive into our methodology, presenting a nuanced understanding of the brain's intricate workings in the context of mood and mental disorders. By leveraging cutting-edge techniques, our study not only enhances our comprehension of these disorders but also paves the way for future investigations, promising new avenues for research and clinical applications.

5.1 Introduction

Schizophrenia (SZ) is a severe psychiatric condition characterized by delusions or hallucinations, loss of motivation, absence of flat affect and disorganized speech or behavior, and cognitive impairment that has a tremendous impact on the patient and society [108], [109]. The development of Magnetic Resonance Imaging (MRI) provided an unprecedented capacity for in vivo studies to explore the anatomy and function of discrete brain regions [110]. Temporal neural activity can be measured by functional magnetic resonance imaging (fMRI) [111], and information regarding the tissue type of the brain can be assessed by structural MRI (sMRI) [112].

Structural MRI studies have shown gray matter volume depletion and cortical mantle thinning across thalamocortical circuitry, together with the prefrontal cortex in patients with chronic SZ, as well as increased ventricular volume [109] [2]. fMRI studies have reported the reduction of functional connectivity of brain areas in SZ [113], [114] and abnormal connectivity was observed between the hippocampus on either side, between distinct cortical areas, the lateral parietal region, and the inferior temporal gyrus [115] [9].

However, the current nosology is not biologically valid, and there are overlaps across mental and mood disorders. The homogeneity within groups of nosology in psychiatry has been investigated in structural MRI shows that current categories do not follow the biological features from fMRI [59], [94]. In this study, we proposed convolutional neural network architectures for functional and structural MRI data to classify psychosis disorder and identify potential label noises from different modalities.

5.2 Materials and Methods

5.2.1 Data Information, Preprocessing, Intrinsic Connectivity Network Extraction

In this work, we analyzed eyes-closed resting-state fMRI and sMRI data collected from the Function Biomedical Informatics Research Network (fBIRN) study [116]. Written informed consent was obtained from all participants. We excluded subjects with the large head motion that led to 160 controls with an age mean of 37.0 ± 10.9 , 45/115 females/males, and 151 subjects with Schizophrenia (SZ) with an age mean of 38.8 ± 11.6 , 36/115 females/males. Age and gender, and site regressed out to avoid confounding effects. Data for six of the seven sites were collected using 3T Siemens Tim Trio System, and 3T General Electric Discovery MR750 scanner was used for one site.

The preprocessing of fMRI data was done using the statistical parametric mapping (SPM12, <http://www.fil.ion.ucl.ac.uk/spm/>) toolbox. The first five scans were eliminated to ensure signal equilibrium and participants' adaptation to the scanner noise. A slice timing and rigid body head motion correction were applied, followed by registration to the standard Montreal Neurological Institute (MNI) space using an echo-planar imaging (EPI) template and were slightly resampled to $3 \times 3 \times 3$ mm³

isotropic voxels. Next, the data were smoothed using a Gaussian kernel with a full width at half maximum (FWHM) of 6 mm. Voxel-level gray matter volume maps and fMRI time series were created independently from the structural data and functional data. The whole-brain GMV measures are represented by a 3D image (size: $121 \times 145 \times 121$) for each subject.

A spatially constrained group independent component analysis approach with the Neuromark pipeline [95] subsequently was applied to the functional data to obtain fifty-three consistent components corresponding to brain areas also known as intrinsic connectivity networks (ICNs) and was categorized into seven functional domains including auditory, cerebellar, cognitive-control, default-mode, sub-cortical, sensorimotor, and visual. We used the fully automated Neuromark pipeline [95], to obtain the same set of ICNs for both control and SZ groups. More details of the Neuromark templates can be found in [95].

5.2.2 *Convolutional Neural Networks (CNNs)*

In this study, we investigated potential neuroanatomical and functional alterations from structural and functional magnetic resonance imaging data. We performed a cross-validated classification task between normal control (NC) and SZ using two deep convolutional neural networks (CNNs).

To capture 3D spatial information from structural data, we used 3D CNN, which comprises four 3D CNN blocks. Each block is made by the stack of 3D convolution, max pooling, batch normalization layers, and dropout layers. The number of filters for convolutional layers are 64, 128, 256, and 512 with kernel size $3 \times 3 \times 3$. Each convolutional

layer is followed by the ReLU activation function and max-pooling layer with a stride of 2. Then, the layer was followed by batch normalization and dropout layers with a dropout rate of 0.3 for regularization and to minimize overfitting. Lastly, an average pooling layer was performed on the last block output, and the flattened output passed into two fully connected layers with the size of 4096 and 512, followed by a dropout layer.

For resting-state fMRI, we used fully convolutional networks to learn features from time courses. The 53 ICNs' time courses passed into the blocks of the 1D convolutional block. Each block starts with 1D convolutional layers with a kernel size of 3 and is followed by a batch normalization layer and ReLU activation layers. The filter size for five 1D convolutional blocks are 2048, 4096, 2048, 1024, and 512. The output passed into average pooling layers and a fully connected layer with a size of 512. The architectures of the models are depicted in Figure 5-1.

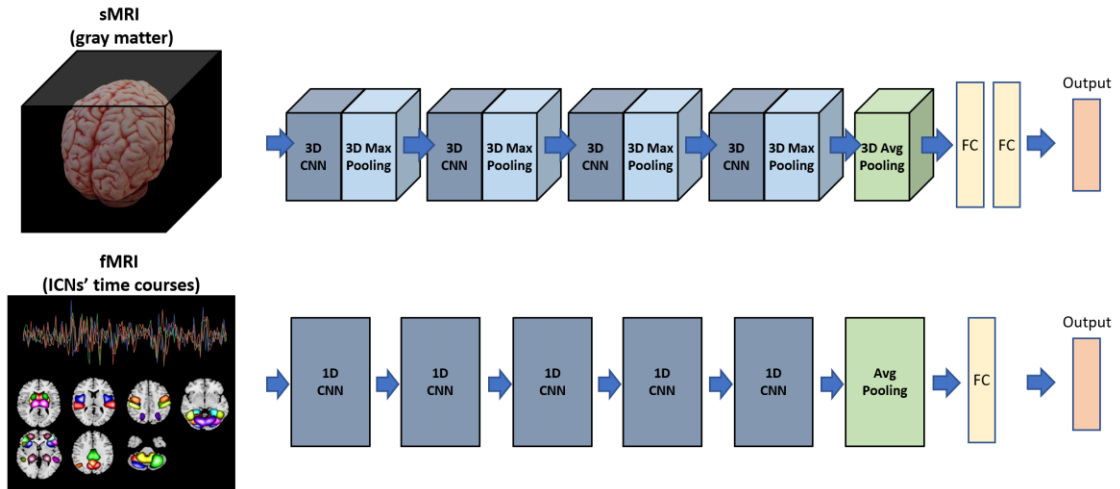


Figure 5-1 – Architecture of our proposed multimodal deep models. Top, 3D convolutional neural network for sMRI data, and bottom, 1D convolutional neural network for fMRI data.

5.2.3 *Classification Tasks and Noise Estimation*

To overcome the unbalance class size, we conducted under sampling method to match the smaller class size and provided balanced subsets. Then, the subset was partitioned into 80% for the train and 20% for the test set using cross-validation. Next, we performed 5-fold stratified cross-validation on the training set. We repeated random cross-validation 5 times. The process repeated till all samples were visited at least in one subset of data. This will help to visit all samples multiple times. We report the models' performance using accuracy, AUC, precision, and recall obtained from the test set.

We defined the noise score by counting the number of times each sample was wrongly classified and divided by the whole number of samples that were classified on the test set. Using repeated cross-validation runs, we obtained a total noise score α for each sample between 0 and 1, with 0 representing the cleanest sample and 1 representing the noisiest sample. To investigate the effect of identified label noise instances, we run the classification tasks again with additional noise information (fMRI-NI and sMRI-NI tasks). We computed sample weights for each instance by

$$w_i = 1 - \alpha_i$$

and train the model with the obtained prior information. It helps mitigate the effects of noisy samples and reduce the complexity of the model by suppressing them and rescaling their loss.

5.2.4 *Groups Comparison*

To evaluate identified noisy subjects, we first calculated static functional connectivity using ICNs' time courses which is the correlation of each pair of ICNs as a

separate measure, and conducted a cell-wise statistical Welch's t-test between different groups. We define noisy samples if the noise criteria are above 0.8. We also defined boundary subjects if their noise score was between 0.3 and 0.8.

5.3 Results

The classification performance of the functional and structural data is reported in Table 5-1. The functional MRI performance revealed 0.87 AUC. The AUC for the structural model is 0.73. Using the noise score as prior, improved the performance of the functional and structural models to 0.89 and 0.79 respectively. The probability distribution obtained from the functional convolutional model is depicted in Figure 5-2. The results show that confusion in the functional model is less than in the structural model.

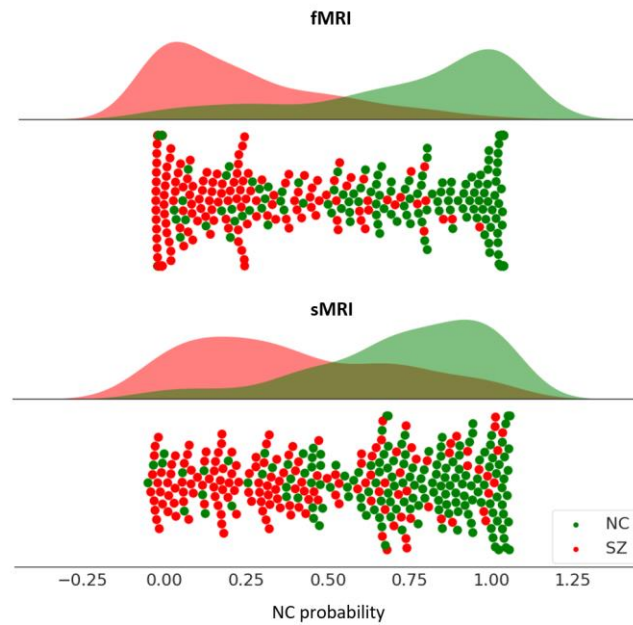


Figure 5-2 – Probability distribution obtain by the model.

To evaluate the noise identification methods and their reliability, we investigate the relationship between the noise scores and static functional connectivity features (sFNC) of

resting-state fMRI data. We identified noisy samples if their noise score, which was obtained as the ratio of the number of misclassifications over the total number of classifications performed on each subject. If the noise score is greater than or equal to 0.8, we consider it as the noisy sample. We calculated the sFNCs for each subject to evaluate our noise criteria. Figure 5-3 shows the average static FC of four groups of clean normal control, noisy normal control, clean Schizophrenia, and noisy Schizophrenia. Figure 5-3 shows that the sFNC pattern of identified noisy samples in normal control is more similar to the sFNC pattern of the clean Schizophrenia group. Similarly, the activation patterns of the noisy Schizophrenia group are more similar to the clean normal control group and they have hyperactivation between the visual, sensorimotor, and subcortical domains as well as higher functional connectivity between the cerebellum and subcortical regions which cannot be detected for the clean Schizophrenia group.

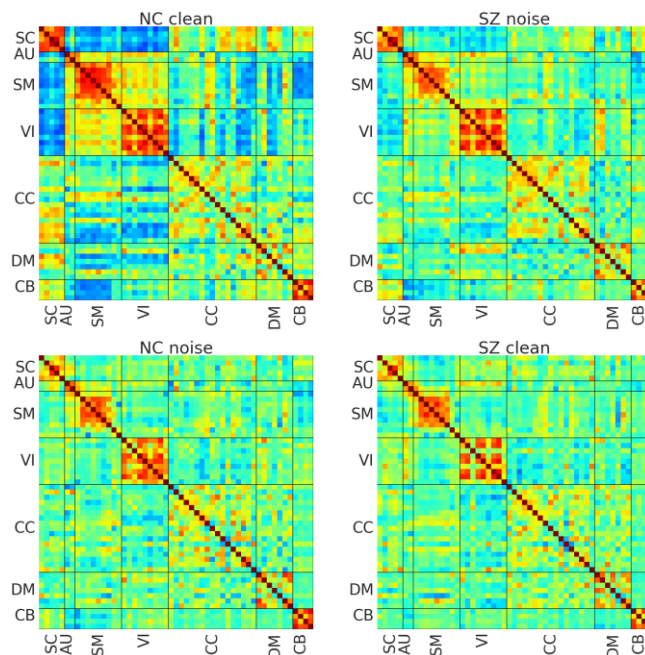


Figure 5-3 – Average of static functional connectivity for different groups. The static FNCs average from clean and noisy subjects of the normal control (NC) group are on

the left panel, and the average of cleansed and noisy samples of the Schizophrenia group are on the right panel.

Table 5-1 – Classification report

	ACC	AUC	Precision	Recall
fMRI	0.81	0.87	0.79	0.80
fMRI-NI	0.83	0.89	0.83	0.83
sMRI	0.67	0.73	0.66	0.66
sMRI-NI	0.72	0.79	0.72	0.73

In addition, we defined 5 different groups based on the estimated noise level and conducted cell-wise group differences between different groups to evaluate the effect of label noise. The statistical test's setup and its results are shown in Figure 5-4. The cell-wise statistical test between the normal control and Schizophrenia groups on original data revealed 715 significant numbers of ICN pairs between all domain regions. The number of significant cells increased to 823, and 913 for clean data with the noise level <0.8 and 0 respectively. The results show that the power of statistical tests improved when the noisy label instances were removed from the tests.

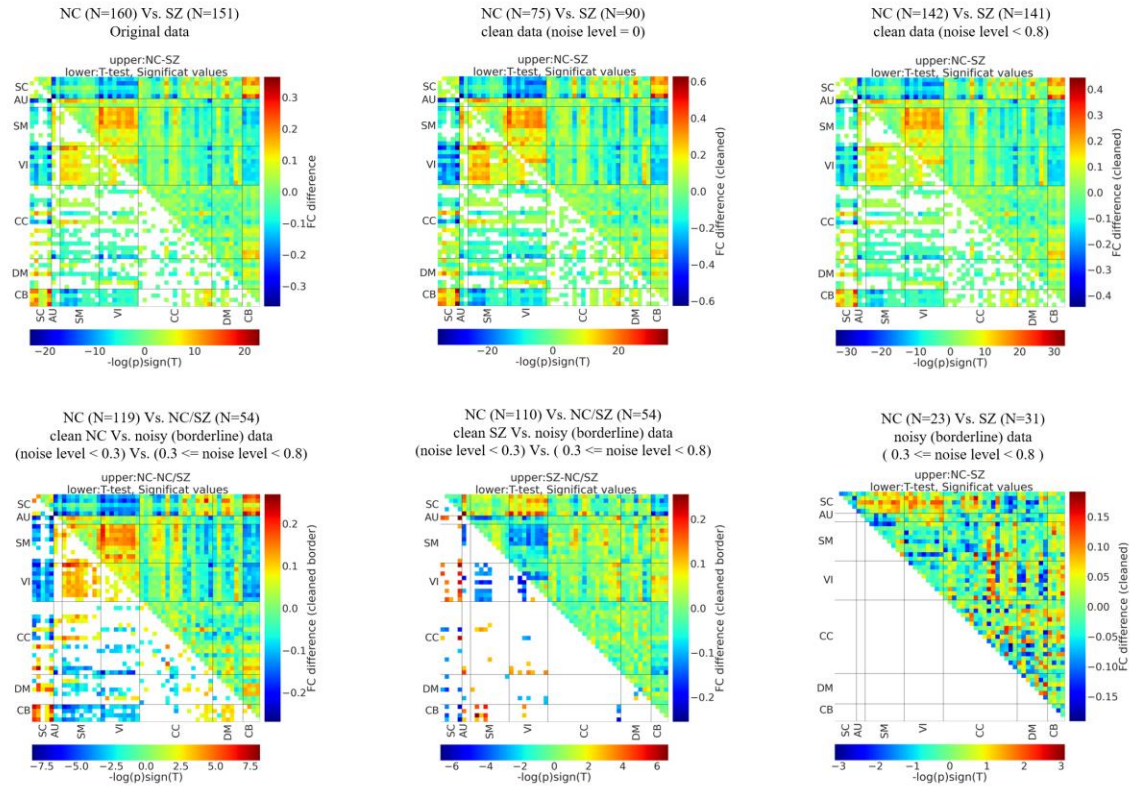


Figure 5-4 – Cell-wise static functional connectivity group difference of different groups. Upper triangle is cell-wise subtraction between functional connectivity of each group and lower triangle shows FDR corrected significant cells obtained from Welch’s t-test at significant level of 0.05

Furthermore, we defined borderline groups as the subjects that have noise scores larger than 0.3 and less than 0.8 and conducted statistical between the clean normal control and clean schizophrenia groups. The test revealed 502 and 115 significant ICN pairs between the borderline group and the clean control groups and between the borderline group and the clean Schizophrenia group. We also performed statistical tests between normal borderline and Schizophrenia groups and the results did not reveal any significant cells among them.

5.4 Discussion

In this study, we proposed convolutional neural networks for functional and structural MRI to classify and identify label noises. Results showed that the resting-state functional MRI data have more related features to distinguish between control groups and Schizophrenia groups in comparison with structural MRI which is consistent with the previous findings.

In addition, we proposed a noise detection approach to identify outliers and noisy data. The label noise score was obtained from different runs using different random subsets of data in cross-validation form. This mitigates the effect of the small sample size for data for deep learning models and the imbalanced sample size. However, it requires more computation recourses. Using sample weight based on the obtained label noise score improved the performance of both models. The sample weights did not contribute to the test set and the reported accuracy. It is only used during the training part to learn representations from most clean data and suppress the noisy data input. Hence, the prior information regarding label noise score and sample weights was not included in the test sets and the reported metrics are based on the original dataset and without the knowledge of label noise in the dataset.

It has been reported that label noise has negative consequences on the statistical test and it might reduce its power. In addition, it might disrupt feature selection and representation learning as important features might not be detected in the presence of label noise. The results show that when noisy label instances were removed from the data, the power of the statistical test improved. The similarity between the normal control and Schizophrenia borderline groups suggests defining subtypes might be beneficial in the diagnosis. In general, the result of group differences among different noise groups reveals

interesting results which suggest defining groups and subtypes based on neurological features might be beneficial in psychiatry nosology.

CHAPTER 6. NEUROIMAGING DATA INFORMED MOOD AND PSYCHOSIS DIAGNOSIS: LABEL NOISE-ROBUST ENSEMBLE DEEP MULTIMODAL FRAMEWORK FOR NEUROIMAGING DATA

In this chapter, we delve into the intricate realm of neuroimaging data, exploring the quest to identify brain-based markers of mental illness. Acknowledging the challenges rooted in symptom-based diagnoses and ambiguous nosological categories, we build upon our prior work that addressed label noise in classification. Our research takes a leap forward by incorporating multimodal neuroimaging data, seamlessly merging functional MRI with structural MRI. To enhance our methodology, we introduce several groundbreaking advancements, leveraging ensemble methods, deep learning techniques, and innovative data fusion approaches.

Central to our study is the utilization of deep convolutional frameworks and bagging techniques for diagnostic classification. Our overarching goal is twofold: to pinpoint potential biomarkers and mitigate the adverse effects of label noise spanning mood and psychosis categories, utilizing both structural and functional MRI data. Rigorous evaluation of our models is carried out through repeated k-fold cross-validation techniques. This comprehensive approach involves training individual base models on distinct data subsets and aggregating independent deep learning models to derive the final classification.

Our findings showcase the superiority of ensemble deep models over individual deep models in both unimodal and multimodal scenarios. Moreover, the multimodal frameworks

consistently outshine their unimodal counterparts in both ensemble and base models, underscoring the power of integrating diverse data sources. Delving deeper, we interpret the results, discerning class-specific learned features crucial to successful diagnoses. Notably, we pinpoint differences in learned features among potentially mislabeled samples, shedding light on nuanced complexities within the data.

Furthermore, our study explores the potential of augmenting existing label categories to account for sample heterogeneity, enriching the depth of our analysis. Through this chapter, we venture into cutting-edge techniques, pushing the boundaries of neuroimaging research and opening new avenues for understanding and diagnosing mental illness.

6.1 Introduction

The rapid advancement of medical imaging techniques, particularly magnetic resonance imaging (MRI), has revolutionized *in vivo* studies aimed at exploring the intricate anatomy and function of discrete brain regions [110], [117]. Structural MRI (sMRI) enables the evaluation of different brain tissue types, while functional magnetic resonance imaging (fMRI) facilitates the measurement of temporal neural activity [111], [112]. Consequently, there has been a growing interest in utilizing the structural and functional features extracted from MRI data as supplementary biological indicators for diagnosing mood and psychosis [1], [118], [119]. The field of brain imaging translational research has embraced advanced machine learning and deep learning techniques to extract knowledge from neuroimaging data and identify patterns associated with various brain and mental disorders [120].

However, the categorization of psychosis and other neuropsychiatric disorders presents a complex challenge, as clear boundaries between different classes are often lacking and overlap between disorders exist [121]–[123]. Consequently, it becomes crucial to consider the limitations inherent in diagnostic validity and reliability, heterogeneity across different psychosis and mood disorders, and the absence of a definitive ground truth during the training and learning process of models [124]. Moreover, distinguishing mislabeled instances from correctly labeled instances is difficult without additional information [38]. To address this issue, an assumption-driven, data-driven approach can be employed to identify potential mislabeled instances and uncover hidden patterns within homogeneous data samples.

Label noise, a common occurrence in classification problems, refers to situations where observed labels are incorrectly assigned to instances, contradicting their true class [38]. The presence of label noise in real-world datasets adversely affects the training process and model performance, leading to potential consequences such as compromised classification accuracy, increased model complexity, altered learning requirements, and impaired feature selection [38]. The identification of relevant and discriminative features becomes challenging in the presence of label noise [38], [125]. In the context of mood and psychotic disorders, the label noise problem becomes even more intricate due to the limitations in diagnostic validity and reliability, the inherent heterogeneity of these disorders, and the absence of a definitive ground truth [76], [126], [127]. The high dimensionality and small sample size further compound the complexity of addressing label noise in these disorders.

Deep learning, as a powerful branch of machine learning, possesses the capacity to learn complex representations from data [101]. However, deep learning models rely heavily on large sample sizes for effective learning [100]. The presence of label noise increases the number of samples required to train a reliable model [44]. Various techniques have been developed to address label noise, including robust losses, ensemble methods, classification voting and filtering approaches, clustering, neighbors-based methods, and semi-supervised approaches [38], [94]. Ensemble methods, in particular, rely on multiple base models rather than a single model [128]. One such ensemble approach is bagging, which aggregates individual base models trained on different subsets of data [128]. Each sampling iteration produces a different model, effectively reducing the impact of mislabeled samples. Bagging has been successfully employed to mitigate the effects of label noise and improve model performance [38], [128]. Furthermore, ensemble methods demonstrate effectiveness in generalization and reducing variance [38], [128].

This study explores the potential of an ensemble deep multimodal framework for the diagnosis of mood and psychosis using structural and functional MRI data, with a particular focus on its robustness in the presence of label noise. Additionally, we investigate the learned features extracted by the models to facilitate better interpretation of the diagnostic outcomes. The organization of this paper is as follows: Section 2 provides an overview of the neuroimaging datasets used, along with details of the preprocessing steps and feature extraction techniques employed. The ensemble deep learning framework and classification structures are also discussed in this section. In Section 3, we present the results obtained by applying our proposed method to different brain imaging modalities. Section 4 discusses

the findings and describes the experimental design employed. Finally, Section 5 concludes the paper, summarizing the key contributions and highlighting areas for future research.

6.2 Materials and Methods

6.2.1 Data information

In this study, we utilize structural and resting-state functional MRI data obtained from three prominent datasets focusing on mood and psychosis research: the first and second waves of the Bipolar-Schizophrenia Network on Intermediate Phenotypes study (B-SNIP-1, B-SNIP-2) [51], [52], [77] and the Function Biomedical Informatics Research Network (FBIRN) study [116]. Prior to participation, written informed consent was obtained from all individuals involved in the respective studies.

The FBIRN study is a collaborative research effort aimed at advancing the understanding of brain function and mental disorders through the use of neuroimaging techniques. The study focuses on investigating the neural underpinnings of psychiatric conditions such as schizophrenia [116]. FBIRN brings together researchers from multiple institutions and employs a multi-site approach, involving the collection of neuroimaging data from various research centers. The study incorporates advanced imaging technologies, such as functional magnetic resonance imaging (fMRI) and structural MRI (sMRI), to examine brain structure, function, and connectivity.

The present study utilized data from the Function Biomedical Informatics Research Network (FBIRN) study, which involved the inclusion of eyes-closed resting-state functional MRI (fMRI) and structural MRI (sMRI) data. A total of 160 control participants

were included in the FBIRN study, with an average age of 37.0 ± 10.9 years. This control group consisted of 45 females and 115 males. Additionally, the study incorporated 151 subjects diagnosed with Schizophrenia (SZ), with an average age of 38.8 ± 11.6 years, including 36 females and 115 males. The acquisition of data in the FBIRN study was carried out using 3T Siemens TIM Trio System scanners across six different sites. Notably, one site employed a 3T General Electric Discovery MR750 scanner [116].

The B-SNIP study was established to investigate genetic and biological factors associated with bipolar and schizophrenia disorders. Multiple assessment measures, including brain imaging, cognitive, electrophysiologic, oculomotor, and genetic measures, were collected from healthy control participants and individuals diagnosed with schizophrenia, schizoaffective disorder, and psychotic bipolar disorder. The aim was to gain a deeper understanding of the underlying biological mechanisms contributing to these disorders and identify potential treatment and prevention targets. Imaging scans were conducted on 3T scanners at multiple sites, with Signa scanners utilized at two sites, an Achieva scanner at one site, and Allegra and Trio scanners at two other sites [53], [61]. Demographic information pertaining to the datasets is provided in Table 6-1.

Table 6-1 – Demographic Table

Dataset	NC		BP		SAD		SZ		Total	
	N (M/F)	Age mean \pm std	N (M/F)	Age mean \pm std	N (M/F)	Age mean \pm std	N (M/F)	Age mean \pm std	N (M/F)	Age mean \pm std
FBIRN	160 (115/45)	37.04 ± 10.86	-	-	-	-	151 (115/36)	38.77 ± 11.63	311 (230/81)	37.88 ± 11.26
B-SNIP-1	234 (100/134)	38.26 ± 12.57	163 (52/111)	36.25 ± 12.76	130 (57/73)	36.48 ± 12.20	182 (125/57)	35.14 ± 12.08	709 (334/375)	36.67 ± 12.46
B-SNIP-2	243 (86/157)	36.23 ± 11.97	98 (40/58)	33.39 ± 10.94	141 (71/70)	40.39 ± 11.08	171 (105/66)	38.66 ± 11.73	653 (302/351)	37.34 ± 11.77
Total	637 (301/336)	37.18 ± 11.95	261 (92/169)	35.18 ± 12.16	271 (128/143)	38.51 ± 11.77	504 (345/159)	37.42 ± 11.93	1673 (866/807)	37.15 ± 11.98

6.2.2 Preprocessing and ICNs extraction

The voxel-level gray matter volume maps and fMRI time series were independently generated from the structural and functional MRI data, respectively. Preprocessing of the structural and functional MRI data was performed using the statistical parametric mapping (SPM12, <http://www.fil.ion.ucl.ac.uk/spm/>) toolbox. For the structural MRI data, the following preprocessing steps were applied: tissue classification, bias correction, and image registration. Subsequently, the structural MRI images were segmented to extract the gray matter brain regions. The resulting 3D gray matter images had a voxel dimension of 121×145×121 for the FBIRN and BSNIP datasets.

To ensure participant adaptation to the scanner and the stability of fMRI signals, the first five scans were discarded. Additionally, a slice timing correction and rigid head-body motion correction were applied to the fMRI data to account for timing differences in slice acquisition and correct for subjects' head motion. The fMRI data were then spatially normalized to the standard Montreal Neurological Institute (MNI) space. Voxel resampling was performed to achieve 3×3×3 mm³ isotropic voxels, and a Gaussian kernel with a full width at half maximum (FWHM) of 6mm was applied for spatial smoothing.

The utilization of multiple sites and scanner models in the FBIRN and B-SNIP studies allows for the inclusion of a diverse range of participants and enhances the generalizability of the findings. However, it is important to note that differences in imaging protocols and scanner characteristics across sites may introduce potential sources of variation that need to be accounted for during data analysis and interpretation. To extract comparable features, a fully automated Neuromark ICA pipeline [95] was employed on the preprocessed data. The utilization of spatially constrained ICNs facilitates the comparison of subjects across different studies, sites, and disorders. Based on prior knowledge, 53

intrinsic connectivity networks (ICNs) were extracted for each subject using the Neuromark_fMRI_1.0 template derived from replicable components across two large independent studies. These ICNs were categorized into seven functional domains, namely auditory, cerebellar, cognitive-control, default-mode, sub-cortical, sensorimotor, and visual. Finally, quality control (QC) measures were implemented to ensure the inclusion of high-quality data [95].

6.2.3 *Classification tasks*

Multiple binary and multiclass classification tasks were performed using a merged dataset obtained from the three aforementioned datasets: FBIRN, B-SNIP-1, and B-SNIP-2. The classification analysis involved a comprehensive set of tasks, including six binary classification tasks representing each possible pair of diagnostic groups, four 3-way classification tasks, and one 4-way classification task encompassing all diagnostic groups. A detailed description of each classification task is provided in Table II, outlining the specific combinations of diagnostic groups and the corresponding classification objectives. This comprehensive approach enables the evaluation and comparison of the ensemble deep multimodal framework's performance across a range of binary and multiclass classification scenarios, providing valuable insights into its effectiveness in mood and psychosis diagnosis.

Table 6-2 – Classification Tasks

Classification Tasks		Number of Subjects														
		FBIRN		B-SNIP1				B-SNIP2				Total				
		NC	SZ	NC	BP	SAD	SZ	NC	BP	SAD	SZ	NC	BP	SAD	SZ	ALL
2-way	NC-BP	160	-	234	163	-	-	243	98	-	-	637	261	-	-	898
	NC-SAD	160	-	234	-	130	-	243	-	141	-	637	-	271	-	908
	NC-SZ	160	151	234	-	-	182	243	-	-	171	637	-	-	504	1141
	SZ-SAD	-	151	-	-	130	182	-	-	141	171	-	-	271	504	775
	SZ-BP	-	151	-	163	-	182	-	98	-	171	-	261	-	504	765
	SAD-BP	-	-	-	163	130	-	-	98	141	-	-	261	271	-	532
3-way	NC-SZ-SAD	160	151	234	-	130	182	243	-	141	171	637	-	271	504	1412
	NC-SAD-BP	160	-	234	163	130	-	243	98	141	-	637	261	271	-	1169
	NC-SZ-BP	160	151	234	163	-	182	243	98	-	171	637	261	-	504	1402
	SZ-SAD-BP	-	151	-	163	130	182	-	98	141	171	-	261	271	504	1036
4-way	NC-SZ-SAD-BP	160	151	234	163	130	182	243	98	141	171	637	261	271	504	1673

6.2.4 Deep learning base models

To extract meaningful features from the structural MRI data, we employed 3D convolutional neural network frameworks. The architecture of our 3D CNN consisted of four convolutional blocks, with filter sizes of 64, 128, 256, and 512 for their respective convolutional layers. Each convolutional layer had a kernel size of 3x3x3. Following each convolutional block, we included a max-pooling layer with a size of 2, batch normalization, and a dropout layer with a rate of 0.5 to mitigate overfitting. The output of the stacked 3D convolutional blocks was flattened and passed through a global average pooling layer, followed by two dense layers with sizes of 4096 and 512, respectively. Finally, a softmax layer was applied to generate the class probabilities. This architecture was specifically chosen to minimize the number of parameters required and enable the learning of distinctive patterns, thus addressing the challenges posed by high-dimensional data and limited sample sizes.

For the representation learning from functional MRI data, we utilized a fully convolutional neural network architecture consisting of five convolutional blocks. Each block included a 1D convolutional layer with a kernel size of 3, followed by a batch normalization layer and ReLU activation. The filter sizes for the convolutional layers within the blocks were defined as 2048, 4096, 2048, 1024, and 512, respectively. The output of the convolutional blocks was then fed into a global average pooling layer, and the final class predictions were obtained from a softmax layer.

In our multimodal analysis, we employed a fusion technique to leverage the complementary information from the structural and functional CNN models. Specifically, we merged the output layers of the two models by averaging their softmax layers and utilizing the mean value to predict the labels. This fusion approach allowed us to combine the distinctive features learned from both modalities, enhancing the model's performance in situations where a single modality may lack sufficient information for accurate predictions.

6.2.5 Ensemble method

We employed a bagging approach to ensemble our ensemble deep method. To introduce data perturbations and train base models using different subsets of data, we implemented a repeated cross-validation procedure consisting of ten repeats of 5-fold cross-validation for all classification tasks. In order to address the challenges posed by imbalanced classes in various tasks, we applied subsampling techniques prior to the cross-validation setup, repeating the process until all samples were visited at least once.

Within each fold, one fold was designated as the unseen test set, while the remaining data was divided into an 80% training set and a 20% validation set to train the models. This approach enabled us to generalize and evaluate our method across the entire dataset, even with limited sample sizes. Furthermore, by repeating the cross-validation process and employing subsampling, we trained diverse base models on different subsets of data and evaluated them using distinct test sets. This allowed us to learn from a wide range of training samples and improve the robustness of our ensemble approach. Finally, for the ensemble method, we aggregated the outputs of independent deep-learning models and computed the average probabilities obtained from the softmax layers. These averaged probabilities were then utilized to predict the labels for the test sets, leveraging the collective knowledge of the ensemble for improved classification accuracy.

Figure 1 provides an overview of the methodology employed in the study, showcasing the steps involved in data subsampling, repeated cross-validation, and the ensemble approach using convolutional neural network models. The diagram illustrates how random subsets of the data are generated for training diverse models, which are then used to predict the labels for unseen test samples. The ensemble approach combines the predictions from multiple models through a voting mechanism, resulting in robust and reliable classification outcomes.

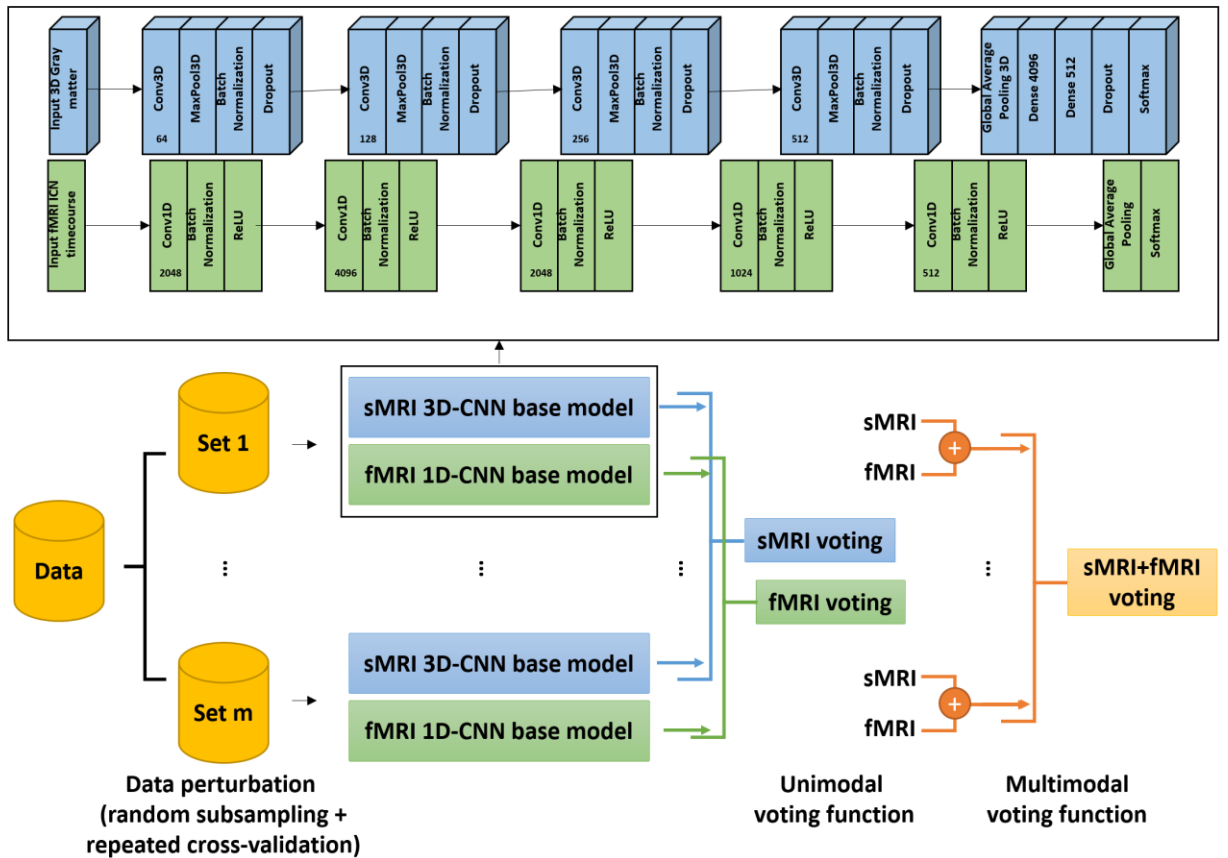


Figure 6-1 – Method Diagram and Convolutional Neural Network Models. The diagram illustrates the methodology employed in this study, incorporating subsampling and a repeated cross-validation approach to enhance the evaluation of each sample using multiple models trained on different subsets of the data. By generating random subsets of the data, we ensure diverse training samples for each model. The predictions from these models are then combined using a voting approach, enabling the determination of final labels for the subjects in both the unimodal and multimodal approaches. This methodology enhances the robustness and generalizability of the classification process by leveraging the collective knowledge of multiple models.

6.2.6 Label noise analysis

Lastly, to gain insights into the classification results and uncover hidden patterns within different classes, we conducted a comprehensive analysis to interpret each task. Our goal was to investigate the heterogeneity present in current psychosis and mental illness categories by examining the differences in features between misclassified and correctly

classified instances for each class. To achieve this, we first examined the probability scores assigned to each class for individual subjects. This allowed us to identify subjects that exhibited high uncertainty or were potentially mislabeled. By assessing the probability scores, we gained valuable information regarding the difficulty of classifying certain subjects and the presence of potential label noise within the dataset.

Next, we performed a thorough investigation of the features that distinguished misclassified instances from correctly classified ones within each class. By analyzing these differences, we aimed to uncover unique characteristics and patterns associated with different diagnostic groups. This analysis provided valuable insights into the underlying heterogeneity within current psychosis and mental illness categories, shedding light on the complexities of these disorders. To further validate our findings, we conducted statistical tests to assess the significance of the observed differences in features between the misclassified and correctly classified groups. This allowed us to determine whether the identified patterns were statistically significant and not merely due to chance.

In the functional MRI modality, functional connectivity features were employed as a means to assess label noises and challenging cases. Static functional connectivity was computed for each subject using the timecourses of intrinsic connectivity networks (ICNs). Subsequently, a statistical t-test was performed to compare functional connectivity between different groups, and significant functional connectivity pairs were reported after applying False Discovery Rate (FDR) correction. This analysis provided insights into the distinctive functional connectivity patterns associated with different diagnostic groups, aiding in the identification of potential sources of misclassification and enhancing the understanding of the underlying neurobiological mechanisms. In a similar vein, within the

structural MRI modality, we performed a statistical t-test on the gray matter images to elucidate group differences. This analysis allowed us to evaluate and compare the distribution of potential label noise and challenging instances with the "clean" groups that were correctly classified by the model. By identifying significant variations in the gray matter distribution, we gained insights into the discriminatory power of the structural features and their relevance to the classification task.

Additionally, we employed the class activation mapping (CAM) technique to further enhance our interpretation of the classification results. CAM allows us to visualize and localize the important regions within the brain that contribute significantly to the classification decision for each specific class. By utilizing CAM, we were able to highlight the regions of interest and provide a visual representation of the neural activation patterns associated with different diagnostic categories. Analyzing the CAM results allowed us to gain valuable insights into the neurobiological underpinnings of these disorders and understand the neural correlates associated with specific diagnostic groups. By identifying the regions that contribute most prominently to the classification decision, we can pinpoint the brain areas that are most relevant in characterizing and differentiating between different mental health conditions.

Overall, the utilization of the CAM technique in our analysis adds another layer of interpretability to our study, allowing us to explore the spatial localization of class-specific neural activations and providing valuable insights into the brain regions that are critical in distinguishing between different diagnostic categories.

6.3 Results

We conducted an extensive investigation into the classification of mood and psychosis disorders using both structural and functional MRI data, employing both unimodal and multimodal approaches. In Table 6-3, we present the classification accuracy and area under the curve (AUC) metrics for eleven different classification tasks using various methods. The results demonstrate the performance of the ensemble deep framework compared to the base models in both unimodal (fMRI and sMRI) and multimodal (fMRI+sMRI) approaches.

Table 6-3 – Classification Results

Tasks		Base model						Ensemble Method					
		fMRI		sMRI		fMRI+sMRI		fMRI		sMRI		fMRI+sMRI	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
2-way tasks	BP-NC	0.76	0.83	0.69	0.74	0.77	0.84	0.86	0.92	0.78	0.85	0.88	0.94
	BP-SAD	0.58	0.60	0.57	0.59	0.59	0.63	0.62	0.64	0.60	0.64	0.64	0.68
	BP-SZ	0.71	0.76	0.72	0.76	0.75	0.81	0.77	0.84	0.76	0.85	0.81	0.89
	SZ-NC	0.76	0.82	0.69	0.76	0.78	0.85	0.79	0.87	0.73	0.81	0.81	0.89
	SZ-SAD	0.68	0.71	0.68	0.74	0.70	0.77	0.73	0.78	0.74	0.81	0.75	0.83
	SAD-NC	0.75	0.82	0.70	0.76	0.77	0.84	0.83	0.90	0.76	0.83	0.83	0.91
3-way tasks	BP-NC-SZ	0.60	0.77	0.55	0.71	0.62	0.80	0.72	0.88	0.66	0.82	0.75	0.90
	BP-NC-SAD	0.53	0.71	0.50	0.68	0.56	0.74	0.67	0.84	0.63	0.80	0.71	0.87
	SZ-SAD-NC	0.57	0.73	0.53	0.71	0.59	0.77	0.66	0.85	0.61	0.80	0.71	0.87
	BP-SZ-SAD	0.48	0.65	0.50	0.69	0.50	0.70	0.55	0.74	0.60	0.78	0.60	0.79
4-way tasks	BP-NC-SZ-SAD	0.45	0.69	0.41	0.67	0.47	0.73	0.58	0.82	0.55	0.78	0.62	0.85

Our results demonstrate that the ensemble deep framework outperforms the base models obtained from cross-validation in all tasks, regardless of whether the unimodal or multimodal approach was employed. Among the binary classification tasks, the ensemble deep multimodal approach achieved the highest accuracy of 0.88 for the BP vs. NC classification task. This was followed by the SAD vs. NC task with an accuracy of 0.83, the SZ vs. NC task with an accuracy of 0.82, and the BP vs. SZ task with an accuracy of

0.81. The BP vs. SAD task and SZ vs. SAD task exhibited relatively lower accuracies among the binary classification tasks, achieving 0.64 and 0.75, respectively.

In the 3-way classification tasks, the ensemble deep multimodal approach consistently outperformed the base models. For instance, in the BP-NC-SZ task, the ensemble method achieved an accuracy of 0.75, whereas the base models using sMRI, fMRI, or their combination achieved lower accuracies ranging from 0.55 to 0.62. This demonstrates the advantage of the ensemble method in capturing the complex relationships among multiple diagnostic groups.

For the 4-way classification task involving all four diagnostic groups (BP, NC, SZ, and SAD), the ensemble deep multimodal approach achieved an accuracy of 0.62, outperforming the base models. Notably, the accuracy for the 3-way classification task BP-SZ-SAD was comparatively lower at 0.60, exhibiting the lowest accuracy among all the classification tasks. However, it is worth noting that this task had the lowest accuracy among all the classification tasks, indicating the difficulty in accurately classifying individuals into the four diagnostic groups.

To provide further insights, Figure 6-2 – The figure presents the confusion matrices resulting from the application of the ensemble deep multimodal approach to classify subjects across 11 different tasks. Each matrix represents the performance of the classification model by displaying the counts or percentages of true positive, true negative, false positive, and false negative predictions for each diagnostic group. The matrices provide a comprehensive visual representation of the model's accuracy and misclassification patterns, offering valuable insights into the categorization outcomes

based on neuroimaging features. presents the confusion matrices obtained from the ensemble deep multimodal approach for the eleven classification tasks. These matrices offer a visual representation of the model's performance, showcasing the distribution of correctly and incorrectly classified instances for each diagnostic group. Our findings underscore the effectiveness of the ensemble deep multimodal approach in accurately classifying mood and psychosis disorders. Our results highlight the potential of utilizing structural and functional MRI data together to improve diagnostic accuracy. These outcomes have significant implications for understanding the heterogeneity within current diagnostic categories and may contribute to the development of more precise and personalized approaches to mental health diagnosis and treatment.

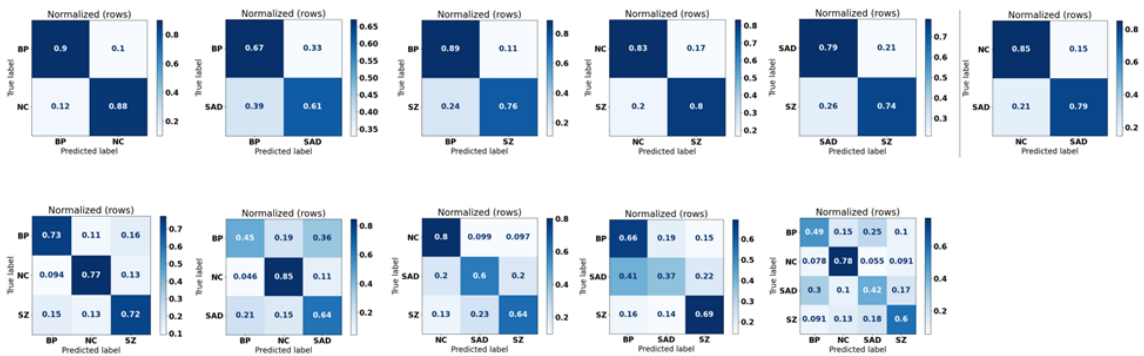


Figure 6-2 – The figure presents the confusion matrices resulting from the application of the ensemble deep multimodal approach to classify subjects across 11 different tasks. Each matrix represents the performance of the classification model by displaying the counts or percentages of true positive, true negative, false positive, and false negative predictions for each diagnostic group. The matrices provide a comprehensive visual representation of the model's accuracy and misclassification patterns, offering valuable insights into the categorization outcomes based on neuroimaging features.

Furthermore, to assess the impact of incorporating different datasets in our study, we conducted a specific classification task involving bipolar disorder and normal control

subjects. In this analysis, we deliberately excluded the FBIRN dataset, which lacks bipolar samples, and trained the model exclusively on the B-SNIP datasets containing bipolar and control groups. This allowed us to investigate the influence of introducing additional data exclusively to one class and examine its effects on the overall performance of the classification model. By systematically examining the performance in this scenario, we gained valuable insights into the role of dataset composition and its implications for accurate classification outcomes.

Table 6-4 and Figure 6-3 provide a comprehensive overview of the results obtained from our analysis. Notably, we observed consistent improvements in model performance across all modalities, both in terms of the base model and the ensemble approach, when incorporating additional data into the analysis. The inclusion of the FBIRN dataset, specifically the control samples, resulted in enhanced classification accuracy for bipolar cases, as demonstrated by the improved performance depicted in the confusion matrix. Moreover, the performance of the model trained exclusively on the B-SNIP datasets, without the inclusion of the FBIRN dataset, was found to be comparable to the model trained using all available datasets. This suggests that the specific composition of the dataset did not significantly impact the model's performance, thereby highlighting the robustness of our approach. These findings underscore the potential benefits of incorporating diverse datasets and shed light on the effectiveness of our methodology in accurately categorizing individuals with bipolar disorder.

Table 6-4 – Classification Performance Using Different Datasets.

Dataset	Base model			Ensemble Method		
	fMRI	sMRI	fMRI+sMRI	fMRI	sMRI	fMRI+sMRI
BSNIP1+BSNIP2+FBIRN	0.76	0.69	0.77	0.86	0.78	0.88
BSNIP1+BSNIP2	0.72	0.64	0.72	0.79	0.72	0.82

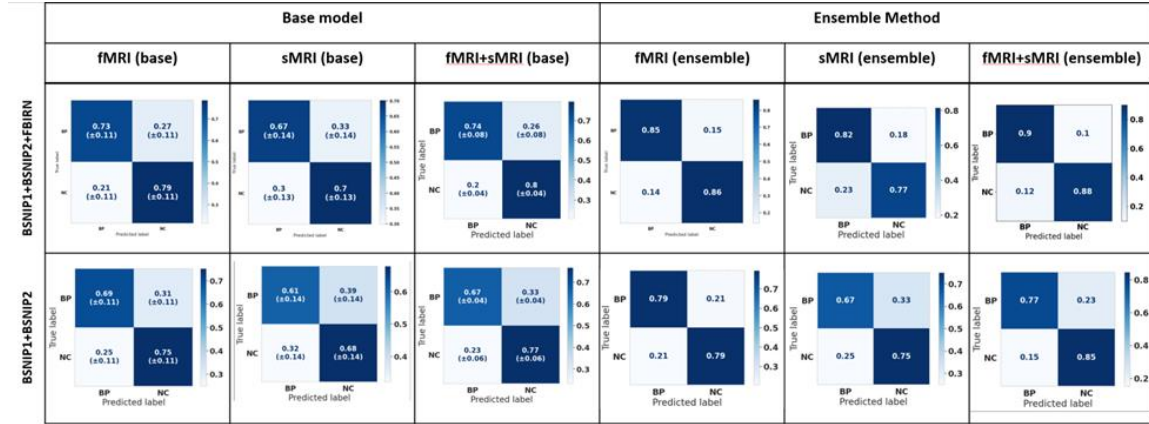


Figure 6-3 – Performance comparison of the models with and without the inclusion of additional data. The results demonstrate improved classification accuracy when more control samples are added using the FBIRN dataset. The model trained solely on the B-SNIP datasets shows comparable performance to the model utilizing all datasets, indicating minimal impact of dataset variation on the model's effectiveness

By leveraging the probabilities obtained from the ensemble approach, we were able to identify potential instances of label noise and challenging cases. Challenging instances were identified when the probabilities assigned to all classes were closely clustered, with differences across classes below 0.1. Similarly, instances that exhibited significantly higher average probabilities for a class other than their assigned label, leading to a confident misclassification, were marked as potential label noise. Therefore, subjects in each class were categorized into one of the following categories: clean, noise, or hard.

Following the categorization of samples for each task, we proceeded to perform statistical tests between different groups based on static functional network connectivity

(FNC) as features from functional MRI, as well as gray matter features from structural MRI. Table 6-5 provides information on group difference analysis using statistical t-tests between classes of different classification tasks based on different types of data. The data is categorized into "All," "Clean," "Noisy," "Suggested," and "Challenging" groups. In the "sMRI (GM)" column, the table presents the number of significant voxels in gray matter (GM) for each class. The total number of voxels in the analysis is 676,567. In the "fMRI (sFNC)" column, the table presents the number of significant functional connectivity (sFNC) pairs for each class. The total number of functional connectivity pairs is 1,378. The table then provides various comparisons between different groups within each classification task. These comparisons include "A vs. B," "A(c) vs. B(c)," "A(c) vs. A(n)," "B(c) vs. B(n)," "A(c) vs. B(n)," and "B(c) vs. A(n)." The letters A and B represent the classes being compared, while (c) and (n) denote clean and noisy label data, respectively. The results obtained from group difference analysis are useful to assess the differences between groups and evaluate the impact of different identified data types on the classification tasks. These findings highlight the differences in significant voxels between different classes and shed light on the distribution of label noise and challenging instances within each class. It suggests that the identified challenging instances may have unique characteristics that make them difficult to classify accurately.

In the binary classification tasks, when considering all data, the comparison between Schizophrenia (SZ) and Normal Control (NC), followed by Schizoaffective (SAD) and Normal Control (NC), demonstrates a higher level of significant difference in the number of significant voxels in structural MRI (sMRI) compared to other tasks. On the other hand, the class comparison between Bipolar (BP) and Normal Control (NC) reveals a lower

number of significant voxels in sMRI. However, when examining functional connectivity (sFNC) data, there is a significant difference between the Bipolar (BP) and Normal Control (NC) groups.

Upon removing identified noisy labeled data and focusing on clean data (correctly classified samples), the analysis shows a significantly higher number of significant voxels in sMRI and FC pairs in sFNC compared to the analysis that includes all data, across all tasks. Moreover, the results indicate that the number of significant features, both in sMRI and fMRI modalities, between noisy samples and their corresponding correctly classified group are higher than the number of significant features between noisy samples and the suggested clean data group. This suggests that the distribution of label noise aligns more closely with the suggested clean data groups.

The aforementioned findings are consistent across both 3-way and 4-way classification tasks. Moreover, there is a notable increase in the number of significant voxels in structural MRI (sMRI) and significant FC pairs in functional MRI (fMRI) when comparing the challenging cases to the Normal Control (NC) class in multi-class tasks. This observation suggests that incorporating more than two diagnoses in the analysis brings attention to the overlapping issues across disorders, making the task of learning distinctive features across classes more complex.

Overall, these findings highlight the distinct patterns of significant voxels and FC pairs across different classification tasks and modalities. They underscore the importance of considering clean data and the impact of label noise on the distribution of significant features in both sMRI and fMRI modalities. The clean data (correctly classified samples)

have a higher number of significant voxels compared to the noisy label data (misclassified samples).

Static functional connectivity pairwise group differences are depicted in Figure 6-4 for the 4-way classification task. The comparison is shown using all available data and only clean data (with label noise removed). It can be observed that the number of significant functional connectivity (FC) pairs increases when using clean data, indicating a reduction in label noise and a clearer differentiation between groups. In addition, Figure 6-5 illustrates the group differences in static functional connectivity (FC) pairs for the 4-way classification task. This figure highlights the comparisons between the noisy data and the corresponding clean group, as well as the noisy data and the suggested clean group.

Table 6-5 – Label Noise Group Difference Analysis

Task Info		sMRI (GM) number of significant voxels (total number of voxels=676567)										fMRI (sFNC) number of significant FC pairs (total number of FC pairs=1378)							
		Al	Cle	Clean	vs.	Clean	vs.	Clean	vs.	Clean	vs.	Al	Cle	Clean	vs.	Clean	vs.	Clean	vs.
k	Task	A	B	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c
		vs))))))))))	vs))))))
-	Task	A	B	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c
		vs))))))))))	vs))))))
w	Task	A	B	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c
		vs))))))))))	vs))))))
a	Task	A	B	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c
		vs))))))))))	vs))))))
y	Task	A	B	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c	A	A(c	A(c	B(c	A(c	B(c	A(c	B(c
		vs))))))))))	vs))))))
2-w a y	BP-NC	B P	N C	17 4	608 9	149	355	208	0	3	773	28 3	440	6	196	0	0	3	2
	BP-SAD	B P	S A	87 9	656 05	293 16	135 41	8	0	370 0	0	1 95	9	4	1	0	1	3	
	BP-SZ	B P	S Z	13 16	167 16	160 9	128	2	143	683	48	20	131	13	7	0	2	0	2
	NC-SAD	N C	S A	57 39	127 566	977 86	198 81	267	0	300 0	420 0	23 5	389	121	1	0	0	1	1
	NC-SZ	N C	S Z	63 80	141 383	560 88	255 81	53	7	347 0	824	34 4	515	61	101	0	2	0	12
	SAD-SZ	S A	S Z	61 7	239	222	217 8	0	75	81	425	8	152	1	189	0	0	0	0
2-w a y	BP-NC-SAD	B P	N C	17 4	685	3	221	0	181	3	401 8	28 3	336	10	80	4	0	1	60
	BP-NC-SZ	B P	S Z	87 13	338 873	635 213	469 127	0	0	3	121 77	1	80	3	3	0	0	1	15
	BP-NC-SZ	N C	S Z	57 38	128 106	758 68	492 7	825	4	401 8	121 77	23 5	296	110	6	0	0	60	15
	BP-NC-SZ	B P	N C	17 4	292 2	505	180 3	800	0	0	32	28 3	351	7	150	0	0	1	159
	BP-NC-SZ	B P	S Z	13 02	873 6	213 5	127	170	1	0	787	20	81	0	6	0	1	1	1
	BP-NC-SZ	N C	S Z	63 79	108 222	205 01	192 4	3	31	32	787	34 4	469	7	44	4	0	159	1
	BP-SAD-SZ	B P	S A	88 4	535 20	112 24	832 5	1	1	373 6	8	1	56	7	1	0	0	1	0
	BP-SAD-SZ	B P	S Z	13 14	270 57	631 7	868	371	0	373 6	350	20	136	4	12	0	0	1	6
	BP-SAD-SZ	S A	S Z	61 6	404 6	0	153 6	100	0	8	350	8	67	0	109	0	0	0	6
	NC-SAD-SZ	N C	S A	57 38	106 100	753 41	100 06	0	356	300 52	17	23 5	218	49	2	0	0	56	0
	NC-SAD-SZ	N C	S Z	63 80	116 821	279 89	192 69	13	7	300 52	185	34 4	511	33	153	10	0	56	17
	NC-SAD-SZ	S A	S Z	61 2	251	11	715	57	0	17	185	8	56	0	108	10	0	0	17
4-w a y	BP-NC-SAD-SZ	B P	N C	17 4	641	4	487	0	279	9	121 91	28 3	274	0	42	0	0	17	222
	BP-NC-SAD-SZ	B P	S A	87 8	350 81	117 23	503 0	0	0	9	273 28	1	31	5	0	1	0	17	0
	BP-NC-SAD-SZ	B P	S Z	12 99	101 36	297	376	7	200	9	472 2	20	119	3	3	0	0	17	17
	BP-NC-SAD-SZ	N C	S A	57 37	101 462	541 44	142 4	376 2	0	121 91	273 28	23 5	124	31	0	0	0	222	0
	BP-NC-SAD-SZ	N C	S Z	63 79	943 53	153 31	963	57	9	121 91	472 2	34 4	412	17	45	0	2	222	17
	BP-NC-SAD-SZ	S A	S Z	61 8	248	45	168 5	0	128	273 28	472 2	8	33	0	92	0	4	0	17

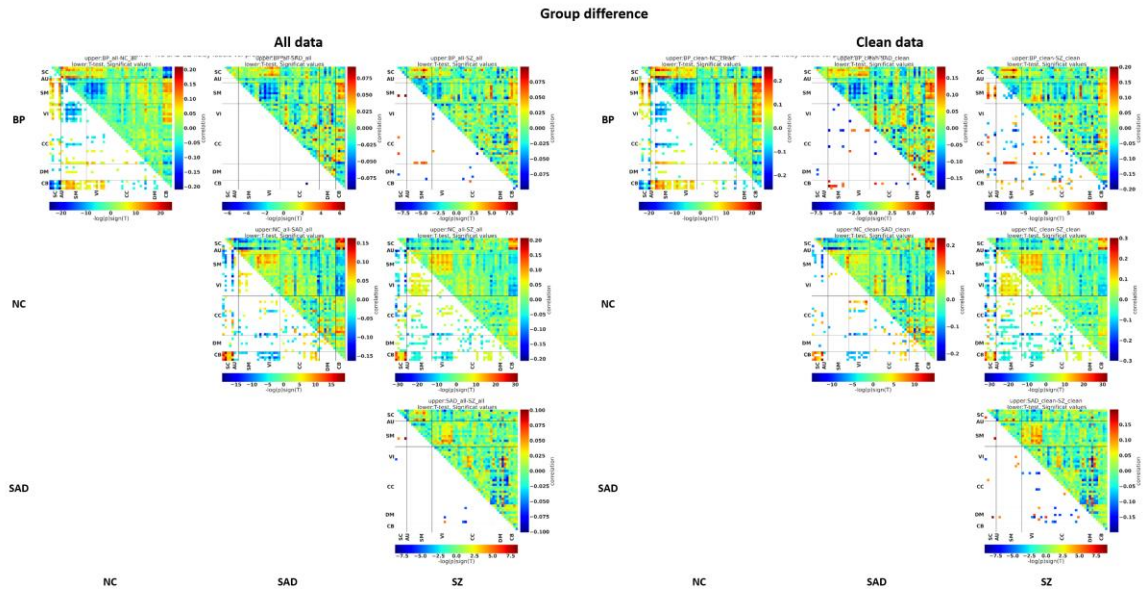


Figure 6-4 – Group differences in static functional connectivity (FC) pairs for the 4-way classification task. Comparison between all data and clean (label noise removed) data shows an increase in the number of significant FC pairs, indicating improved discrimination between groups.

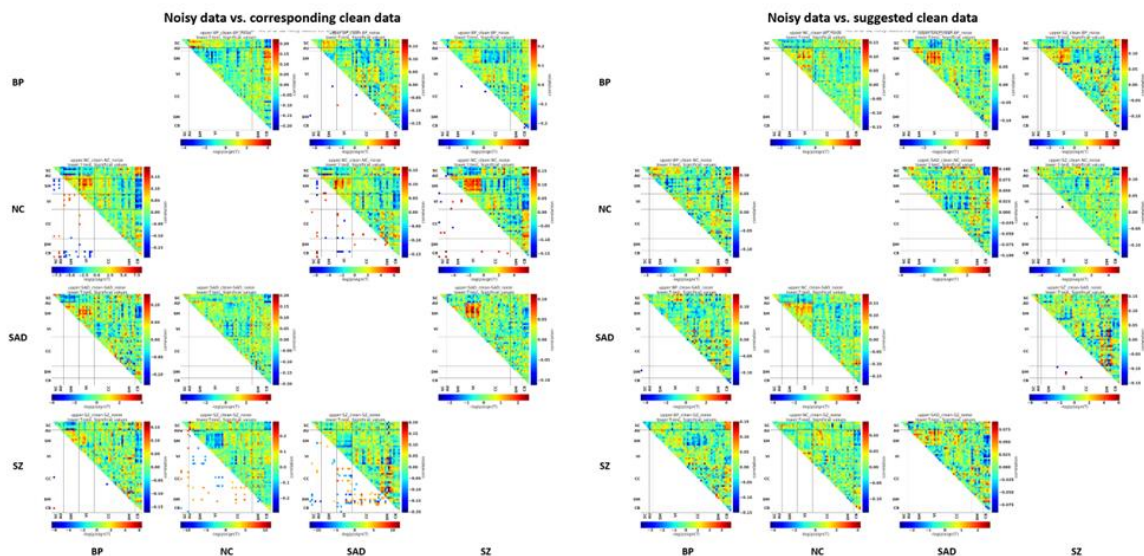


Figure 6-5 – Group differences in static functional connectivity (FC) pairs for the 4-way classification task. Comparison between noisy data and the corresponding clean group, as well as noisy data and the suggested clean group, reveals distinctive patterns of FC differences, highlighting the presence of label noise and the potential for improving classification accuracy.

6.4 Discussion

The results highlight the benefit of the ensemble deep multimodal approach over the base models in terms of classification accuracy and AUC metrics across various tasks. The multimodal fusion of fMRI and sMRI data enables the model to leverage complementary information from different modalities, leading to improved performance in mood and psychosis categorization. The results of the classification analysis provide insights into the potential utility of these modalities for categorization in the Diagnostic and Statistical Manual of Mental Disorders (DSM) in the future. While the classification accuracies varied across different tasks and modalities, the results suggest that neuroimaging features may provide valuable information to complement the existing diagnostic criteria outlined in the DSM.

The higher accuracies achieved in some of the binary classification tasks, such as BP vs. NC, SZ vs. NC, and BP vs. SZ, suggest that neuroimaging features have the potential to aid in the differentiation between specific diagnostic groups. However, for the schizoaffective disorder group, the results suggest that it may not be a useful category in the context of mood and psychosis classification, especially where bipolar disorders are involved. The accuracies achieved in the classification tasks involving schizoaffective and bipolar were generally lower compared to other tasks. For example, in the bipolar vs. schizoaffective, the lower accuracy of the ensemble deep multimodal approach indicates that distinguishing between bipolar disorder and schizoaffective is challenging. In general, these lower accuracies in the classification tasks involving schizoaffective indicate that the features captured by the neuroimaging modalities (sMRI and fMRI) may not effectively differentiate individuals with schizoaffective disorder from those with bipolar disorder or

schizophrenia. The neuroimaging features may not capture distinct patterns that are specific to the schizoaffective category, or the neuroimaging features are influenced by overlapping characteristics of bipolar disorder and schizophrenia, making it more challenging to accurately classify individuals into the schizoaffective group. This suggests that the current diagnostic criteria for schizoaffective may need further refinement or consideration of additional clinical or biological markers to improve its categorization. Furthermore, it should be noted that the sample sizes for the bipolar and schizoaffective groups were comparatively smaller than those of the other groups, which adds an additional layer of complexity in distinguishing between these two diagnostic categories.

By undertaking this comprehensive investigation of the classification results and analyzing the differences between misclassified and correctly classified instances, we aimed to deepen our understanding of the challenges and complexities associated with diagnosing mood and psychosis disorders. The presence of label noise and hard samples in all tasks highlights the challenges inherent in accurately categorizing individuals with mood and psychosis disorders. Label noise refers to mislabeled or ambiguous instances within the dataset, which can adversely affect classification performance. Our analysis reveals that the identified noisy labels and hard samples differ in their characteristics across the tasks, emphasizing the need for careful data curation and the development of robust classification models that can handle such challenges. Furthermore, the heterogeneity within diagnostic categories poses a significant challenge in classification tasks. The overlapping characteristics of bipolar disorder, schizoaffective disorder, and schizophrenia make it more difficult to accurately differentiate individuals into distinct diagnostic groups. This suggests that the current diagnostic criteria, which rely on clinical symptoms and self-

reported experiences, may not fully capture the underlying neurobiological heterogeneity within these disorders.

We conducted multi-study psychosis datasets in this study for better generalization and validation. However, it is important to acknowledge certain limitations of our study. First, the generalizability of our findings may be constrained by the specific datasets and population characteristics included in our analysis. Replication of our results using larger and more diverse samples is necessary to validate the robustness and generalizability of our classification models. Additionally, the integration of other modalities, such as genetic or proteomic data, could further enhance our understanding of the complex etiology of mood and psychosis disorders. Continued research efforts, involving larger and more diverse samples, as well as the integration of additional clinical or biological markers, such as genetic information or cognitive measures, of other modalities and data sources, will be crucial in refining the diagnostic categories and potentially improving the precision and accuracy of psychiatric classification in the future.

In addition to the aforementioned key points, our study sheds light on several important aspects regarding the classification of mood and psychosis disorders. One notable finding is the variation in classification accuracies across different tasks and modalities. This suggests that the effectiveness of neuroimaging features in distinguishing between diagnostic groups may depend on the specific task and the modalities utilized. Understanding these variations can guide the selection of appropriate modalities and feature combinations for future classification studies.

6.5 Conclusion

In conclusion, our study investigated the classification performance of a deep multimodal approach using structural MRI (sMRI) and functional MRI (fMRI) features for the identification of different psychiatric disorders. We demonstrated that incorporating both sMRI and fMRI modalities improved the classification accuracy compared to using each modality individually. In addition, we categorized subjects into clean, noisy, and challenging groups, highlighting the presence of label noise and difficult-to-classify instances. The ensemble approach showed promising results in the presence of label noise. Furthermore, our findings revealed that adding more data to the analysis, particularly clean data, enhanced the performance of the classification models across all modalities. The statistical analysis of group differences in sMRI and fMRI features indicated significant variations between different diagnostic classes. It is important to acknowledge the limitations inherent in diagnostic validity and reliability, as well as the heterogeneity across different psychosis and mood disorders. The absence of a definitive ground truth during the model training process underscores the challenges associated with developing accurate classification models in the field of psychiatric research. Overall, our study provides insights into the potential of deep multimodal approaches in psychiatric classification and highlights the importance of considering the limitations and complexities associated with diagnosing and classifying mental disorders. Future research should continue to explore innovative approaches to improve the robustness and generalizability of psychiatric classification models, taking into account the diverse characteristics and diagnostic challenges within this domain. The results of this study provide a foundation for future research and support the potential inclusion of neuroimaging features as an adjunctive tool

in the DSM for enhanced categorization of mood and psychosis disorders. By incorporating neuroimaging data alongside clinical assessments, it may be possible to develop more precise and objective diagnostic criteria that reflect the underlying neurobiological mechanisms and patterns associated with these disorders. The results of this study provide a foundation for future research and support the potential inclusion of neuroimaging features as an adjunctive tool in the DSM for enhanced categorization of mood and psychosis disorders.

CHAPTER 7. CONCLUSION

In this comprehensive exploration of neuroimaging data, our research endeavors have significantly contributed to the ongoing efforts in psychiatric classification. By harnessing advanced techniques and innovative methodologies, our studies have shed light on the intricate landscape of mental disorders, particularly focusing on mood and psychosis disorders. The convergence of structural MRI (sMRI) and functional MRI (fMRI) modalities has emerged as a pivotal avenue, demonstrating that the integration of these data streams remarkably enhances classification accuracy. This multimodal approach not only refines our understanding of psychiatric disorders but also paves the way for more effective diagnostic strategies.

One of the key accomplishments of our research lies in the identification and management of label noise prevalent in diagnostic classifications. Through a novel iterative classification voting filtering technique, we successfully tackled the challenge of incorrect diagnostic labels. This approach, informed by high-dimensional biological data, enabled us to assign new labels to inconsistently diagnosed individuals, thereby enhancing the reliability of our models. Moreover, our findings unveiled the nuances within the diagnostic categories, emphasizing the need for a nuanced, semi-blind categorization approach that incorporates both clinical expertise and comprehensive data analysis.

Additionally, our study delved into the realm of dynamic functional network connectivity derived from resting-state fMRI data. We meticulously analyzed these connectivity patterns using state-of-the-art machine learning techniques, comparing traditional symptom-based DSM-IV categorization with a biologically informed Biotype

categorization approach. Our results offered valuable insights into the potential of both categorization methods, highlighting their strengths and limitations. Notably, the challenge of overlap among diverse patient groups and controls in multi-class problems emerged as a significant hurdle. Despite these challenges, the integration of biologically based measures, such as neuroimaging data, stands out as a promising direction for the future of nosology in mood and psychosis disorders. The bridge between clinical observations and high-dimensional data holds immense potential, shaping the landscape of psychiatric research.

Furthermore, our research unveiled the power of deep multimodal approaches in psychiatric classification. By harnessing the synergy between sMRI and fMRI features, we achieved enhanced classification accuracy, surpassing the outcomes achievable with individual modalities. The meticulous categorization of subjects into clean, noisy, and challenging groups provided a nuanced understanding of label noise and complex instances. Our ensemble approach exhibited remarkable resilience in the face of label noise, underlining the robustness of our methodology.

However, it is crucial to acknowledge the inherent challenges in psychiatric research, particularly concerning diagnostic validity, reliability, and the heterogeneity prevalent across different disorders. The absence of a definitive ground truth during model training underscores the complexities associated with developing accurate classification models. Our findings, while illuminating, emphasize the need for ongoing research to refine and generalize psychiatric classification models. As we move forward, it is imperative to consider the multifaceted nature of mental disorders, embracing innovative approaches that amalgamate clinical insights with cutting-edge technologies.

In summary, our research endeavors have significantly advanced the field of psychiatric classification, offering a glimpse into the potential of deep multimodal approaches. By addressing label noise, integrating diverse data streams, and navigating the complexities of mental disorders, our studies have set the stage for future research. The integration of neuroimaging features into diagnostic frameworks holds the promise of revolutionizing the categorization of mood and psychosis disorders. As we continue to unravel the mysteries of the human brain, our findings provide a foundation for future endeavors, steering us toward more precise, objective, and comprehensive diagnostic criteria that mirror the underlying neurobiological intricacies of these disorders.

REFERENCES

- [1] B. Rashid and V. Calhoun, “Towards a brain-based predictive of mental illness,” *Hum. Brain Mapp.*, vol. 41, no. 12, pp. 3468–3535, 2020, doi: 10.1002/hbm.25013.
- [2] P. Falkai, A. Schmitt, and N. Andreasen, “Forty years of structural brain imaging in mental disorders: is it clinically useful or not?,” *Dialogues Clin. Neurosci.*, vol. 20, no. 3, pp. 179–186, 2018, doi: 10.31887/dcms.2018.20.3/pfalkai.
- [3] S. Klöppel *et al.*, “Automatic classification of MR scans in Alzheimer’s disease,” *Brain*, vol. 131, no. 3, pp. 681–689, 2008, doi: 10.1093/brain/awm319.
- [4] S. Teipel *et al.*, “Multimodal imaging in Alzheimer’s disease: validity and usefulness for early detection,” *Lancet Neurol.*, vol. 14, no. 10, pp. 1037–1053, 2015, doi: 10.1016/s1474-4422(15)00093-9.
- [5] P. Vemuri *et al.*, “Antemortem differential diagnosis of dementia pathology using structural MRI: Differential-STAND,” *NeuroImage*, vol. 55, no. 2, pp. 522–531, 2011, doi: 10.1016/j.neuroimage.2010.12.073.
- [6] S. J. Teipel, J. Kurth, B. Krause, M. J. Grothe, and for the A. D. N. Initiative, “The relative importance of imaging markers for the prediction of Alzheimer’s disease dementia in mild cognitive impairment — Beyond classical regression,” *NeuroImage: Clin.*, vol. 8, pp. 583–593, 2015, doi: 10.1016/j.nicl.2015.05.006.
- [7] H. A. Helaly, M. Badawy, and A. Y. Haikal, “Deep Learning Approach for Early Detection of Alzheimer’s Disease,” *Cogn. Comput.*, vol. 14, no. 5, pp. 1711–1727, 2022, doi: 10.1007/s12559-021-09946-2.
- [8] E. Hosseini-Asl, R. Keynton, and A. El-Baz, “Alzheimer’s Disease Diagnostics by Adaptation of 3D Convolutional Network,” *2016 IEEE Int. Conf. Image Process. (ICIP)*, pp. 126–130, 2016, doi: 10.1109/icip.2016.7532332.
- [9] A. Nunes *et al.*, “Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working

- Group,” *Mol Psychiatr*, vol. 25, no. 9, pp. 2130–2143, 2020, doi: 10.1038/s41380-018-0228-9.
- [10] E. Cheon *et al.*, “Cross disorder comparisons of brain structure in schizophrenia, bipolar disorder, major depressive disorder, and 22q11.2 deletion syndrome: A review of ENIGMA findings,” *Psychiatry Clin. Neurosci.*, vol. 76, no. 5, pp. 140–161, 2022, doi: 10.1111/pcn.13337.
- [11] V. Rocha-Rego, J. Jogia, A. F. Marquand, J. Mourao-Miranda, A. Simmons, and S. Frangou, “Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach,” *Psychol. Med.*, vol. 44, no. 3, pp. 519–532, 2014, doi: 10.1017/s0033291713001013.
- [12] H. G. Schnack *et al.*, “Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects,” *NeuroImage*, vol. 84, pp. 299–306, 2014, doi: 10.1016/j.neuroimage.2013.08.053.
- [13] T. G. M. van Erp *et al.*, “Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium,” *Mol. Psychiatry*, vol. 21, no. 4, pp. 547–553, 2016, doi: 10.1038/mp.2015.63.
- [14] I. C. Gould, A. M. Shepherd, K. R. Laurens, M. J. Cairns, V. J. Carr, and M. J. Green, “Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach,” *NeuroImage: Clin.*, vol. 6, pp. 229–236, 2014, doi: 10.1016/j.nicl.2014.09.009.
- [15] S. Guo, L. Palaniyappan, P. F. Liddle, and J. Feng, “Dynamic cerebral reorganization in the pathophysiology of schizophrenia: a MRI-derived cortical thickness study,” *Psychol. Med.*, vol. 46, no. 10, pp. 2201–2214, 2016, doi: 10.1017/s0033291716000994.
- [16] P. Yushkevich, A. Dubb, Z. Xie, R. Gur, R. Gur, and J. Gee, “Regional Structural Characterization of the Brain of Schizophrenia Patients1,” *Acad. Radiol.*, vol. 12, no. 10, pp. 1250–1261, 2005, doi: 10.1016/j.acra.2005.06.014.
- [17] M. Nieuwenhuis, N. E. M. van Haren, H. E. H. Pol, W. Cahn, R. S. Kahn, and H. G. Schnack, “Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples,” *NeuroImage*, vol. 61, no. 3, pp. 606–612, 2012, doi: 10.1016/j.neuroimage.2012.03.079.

- [18] D. P. Hibar *et al.*, “Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group,” *Mol Psychiatr*, vol. 23, no. 4, pp. 932–942, 2018, doi: 10.1038/mp.2017.73.
- [19] R. C. Craddock, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, “Disease state prediction from resting state functional connectivity,” *Magn. Reson. Med.*, vol. 62, no. 6, pp. 1619–1628, 2009, doi: 10.1002/mrm.22159.
- [20] Y. Yu, H. Shen, L.-L. Zeng, Q. Ma, and D. Hu, “Convergent and Divergent Functional Connectivity Patterns in Schizophrenia and Depression,” *PLoS ONE*, vol. 8, no. 7, p. e68250, 2013, doi: 10.1371/journal.pone.0068250.
- [21] V. D. Calhoun and T. Adalı, “Feature-Based Fusion of Medical Imaging Data,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 711–720, 2009, doi: 10.1109/titb.2008.923773.
- [22] V. D. Calhoun and J. Sui, “Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness,” *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging*, vol. 1, no. 3, pp. 230–244, 2016, doi: 10.1016/j.bpsc.2015.12.005.
- [23] S. Liu *et al.*, “Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders,” *Brain Informatics*, vol. 2, no. 3, pp. 167–180, 2015, doi: 10.1007/s40708-015-0019-x.
- [24] V. D. Calhoun, T. Adalı, K. A. Kiehl, R. Astur, J. J. Pekar, and G. D. Pearlson, “A method for multitask fMRI data fusion applied to schizophrenia,” *Hum. Brain Mapp.*, vol. 27, no. 7, pp. 598–610, 2006, doi: 10.1002/hbm.20204.
- [25] J. Sui, T. Adalı, Q. Yu, J. Chen, and V. D. Calhoun, “A review of multivariate methods for multimodal fusion of brain imaging data,” *J. Neurosci. Methods*, vol. 204, no. 1, pp. 68–81, 2012, doi: 10.1016/j.jneumeth.2011.10.031.
- [26] J. Ford, L. Shen, F. Makedon, L. A. Flashman, and A. J. Savkin, “A Combined Structural-Functional Classification of Schizophrenia using Hippocampal Volume plus fMRI Activation,” *Proc. Second Jt. 24th Annu. Conf. Annu. Fall Meet. Biomed. Eng. Soc. Eng. Med. Biol.*, vol. 1, pp. 48–49, 2002, doi: 10.1109/iembs.2002.1134381.

- [27] C. Cabral *et al.*, “Classifying Schizophrenia Using Multimodal Multivariate Pattern Recognition Analysis: Evaluating the Impact of Individual Clinical Profiles on the Neurodiagnostic Performance,” *Schizophr. Bull.*, vol. 42, no. suppl_1, pp. S110–S117, 2016, doi: 10.1093/schbul/sbw053.
- [28] M. N. I. Qureshi, J. Oh, D. Cho, H. J. Jo, and B. Lee, “Multimodal Discrimination of Schizophrenia Using Hybrid Weighted Feature Concatenation of Brain Functional Connectivity and Anatomical Features with an Extreme Learning Machine,” *Front Neuroinform*, vol. 11, p. 59, 2017, doi: 10.3389/fninf.2017.00059.
- [29] H. Yang, H. He, and J. Zhong, “Multimodal MRI characterisation of schizophrenia: a discriminative analysis,” *Lancet*, vol. 388, p. S36, 2016, doi: 10.1016/s0140-6736(16)31963-8.
- [30] M. A. Rahaman *et al.*, “Deep multimodal predictome for studying mental disorders,” *Hum. Brain Mapp.*, vol. 44, no. 2, pp. 509–522, 2023, doi: 10.1002/hbm.26077.
- [31] J. Oh, B.-L. Oh, K.-U. Lee, J.-H. Chae, and K. Yun, “Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm,” *Front. Psychiatry*, vol. 11, p. 16, 2020, doi: 10.3389/fpsy.2020.00016.
- [32] A. Abrol *et al.*, “Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning,” *Nat Commun*, vol. 12, no. 1, p. 353, 2021, doi: 10.1038/s41467-020-20655-6.
- [33] Z. Wang *et al.*, “Large-Scale Fusion of Gray Matter and Resting-State Functional MRI Reveals Common and Distinct Biological Markers across the Psychosis Spectrum in the B-SNIP Cohort,” *Front. Psychiatry*, vol. 6, p. 174, 2015, doi: 10.3389/fpsy.2015.00174.
- [34] G. D. Pearlson, B. A. Clementz, J. A. Sweeney, M. S. Keshavan, and C. A. Tamminga, “Does Biology Transcend the Symptom-based Boundaries of Psychosis?,” *Psychiatr. Clin. North Am.*, vol. 39, no. 2, pp. 165–174, 2016, doi: 10.1016/j.psc.2016.01.001.
- [35] T. R. Insel and B. N. Cuthbert, “Brain disorders? Precisely,” *Science*, vol. 348, no. 6234, pp. 499–500, 2015, doi: 10.1126/science.aab2358.
- [36] C. P. Langlotz *et al.*, “A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy

- Workshop,” *Radiology*, vol. 291, no. 3, p. 190613, 2019, doi: 10.1148/radiol.2019190613.
- [37] T. Insel *et al.*, “Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders,” *Am. J. Psychiatry*, vol. 167, no. 7, pp. 748–751, 2010, doi: 10.1176/appi.ajp.2010.09091379.
- [38] B. Fréney and M. Verleysen, “Classification in the Presence of Label Noise: A Survey,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014, doi: 10.1109/tnnls.2013.2292894.
- [39] I. Bross, “Misclassification in 2 x 2 tables,” *International Biometric Society*, vol. 10, no. 4, pp. 478–486, 1954, doi: 10.2307/3001619.
- [40] A. Hadgu, “The discrepancy in discrepant analysis,” *Lancet*, vol. 348, no. 9027, pp. 592–593, 1996, doi: 10.1016/s0140-6736(96)05122-7.
- [41] B. Fréney and A. Kabán, “A Comprehensive Introduction to Label Noise,” in *ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges (Belgium), 2014. [Online]. Available: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c44f388832d6f309b1bb9ccdeddee491f195e6cd>
- [42] R. J. Hickey, “Noise modelling and evaluating learning from examples,” *Artif. Intell.*, vol. 82, no. 1–2, pp. 157–179, 1996, doi: 10.1016/0004-3702(94)00094-8.
- [43] X. Zhu and X. Wu, “Class Noise vs. Attribute Noise: A Quantitative Study,” *Artif. Intell. Rev.*, vol. 22, no. 3, pp. 177–210, 2004, doi: 10.1007/s10462-004-0751-8.
- [44] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, “Deep Learning is Robust to Massive Label Noise,” *Arxiv*, 2017, doi: 10.48550/arxiv.1705.10694.
- [45] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting Unreasonable Effectiveness of Data in Deep Learning ERA,” *2017 IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 843–852, 2017, doi: 10.1109/iccv.2017.97.
- [46] A. Abrol, H. Rokham, and V. D. Calhoun, “Diagnostic and Prognostic Classification of Brain Disorders Using Residual Learning on Structural MRI Data*,” in *2019 41st*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 4084–4088. doi: 10.1109/embc.2019.8857902.

- [47] B. L. Amann *et al.*, “Brain structural changes in schizoaffective disorder compared to schizophrenia and bipolar disorder,” *Acta Psychiatr. Scand.*, vol. 133, no. 1, pp. 23–33, 2016, doi: 10.1111/acps.12440.
- [48] H. Falakshahi *et al.*, “Meta-Modal Information Flow: A Method for Capturing Multimodal Modular Disconnectivity in Schizophrenia,” *Ieee T Bio-med Eng*, vol. 67, no. 9, pp. 2572–2584, 2019, doi: 10.1109/tbme.2020.2964724.
- [49] A. Ortiz, J. M. Górriz, J. Ramírez, F. J. Martínez-Murcia, and for the A. D. N. Initiative, “LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimer’s disease,” *Pattern Recognit. Lett.*, vol. 34, no. 14, pp. 1725–1733, 2013, doi: 10.1016/j.patrec.2013.04.014.
- [50] M. S. Keshavan *et al.*, “A dimensional approach to the psychosis spectrum between bipolar disorder and schizophrenia: The Schizo-Bipolar Scale,” *Schizophr. Res.*, vol. 133, no. 1–3, pp. 250–254, 2011, doi: 10.1016/j.schres.2011.09.005.
- [51] C. A. Tamminga *et al.*, “Clinical Phenotypes of Psychosis in the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP),” *Am J Psychiat*, vol. 170, no. 11, pp. 1263–1274, 2013, doi: 10.1176/appi.ajp.2013.12101339.
- [52] B. A. Clementz *et al.*, “Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers,” *Am J Psychiat*, vol. 173, no. 4, pp. 373–384, 2016, doi: 10.1176/appi.ajp.2015.14091200.
- [53] E. I. Ivleva *et al.*, “Brain Structure Biomarkers in the Psychosis Biotypes: Findings From the Bipolar-Schizophrenia Network for Intermediate Phenotypes,” *Biol Psychiat*, vol. 82, no. 1, pp. 26–39, 2017, doi: 10.1016/j.biopsych.2016.08.030.
- [54] C. A. Tamminga, G. Pearlson, M. Keshavan, J. Sweeney, B. Clementz, and G. Thaker, “Bipolar and Schizophrenia Network for Intermediate Phenotypes: Outcomes Across the Psychosis Continuum,” *Schizophrenia Bull*, vol. 40, no. Suppl_2, pp. S131–S137, 2014, doi: 10.1093/schbul/sbt179.
- [55] J. Ashburner and K. J. Friston, “Unified segmentation,” *NeuroImage*, vol. 26, no. 3, pp. 839–851, 2005, doi: 10.1016/j.neuroimage.2005.02.018.

- [56] F. Pedregosa *et al.*, “Scikit-Learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, 2011.
- [57] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987, doi: 10.1016/0169-7439(87)80084-9.
- [58] B. Mwangi, T. S. Tian, and J. C. Soares, “A Review of Feature Reduction Techniques in Neuroimaging,” *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014, doi: 10.1007/s12021-013-9204-3.
- [59] H. Rokham, H. Falakshahi, and V. D. Calhoun, “A data-driven approach for stratifying psychotic and mood disorders subjects using structural magnitude resonance imaging data,” in *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314. 2020, pp. 113142V–113142V–10. doi: 10.1117/12.2549680.
- [60] J. R. Quinlan, “Induction of decision trees,” *Mach Learn*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/bf00116251.
- [61] E. I. Ivleva *et al.*, “Gray Matter Volume as an Intermediate Phenotype for Psychosis: Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP),” *Am J Psychiat*, vol. 170, no. 11, pp. 1285–1296, 2013, doi: 10.1176/appi.ajp.2013.13010126.
- [62] C. R. Jack *et al.*, “The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods,” *J. Magn. Reson. Imaging*, vol. 27, no. 4, pp. 685–691, 2008, doi: 10.1002/jmri.21049.
- [63] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. in Princeton Legacy Library. Princeton University Press, 1961.
- [64] K. Allsopp, J. Read, R. Corcoran, and P. Kinderman, “Heterogeneity in psychiatric diagnostic classification,” *Psychiatry Res.*, vol. 279, pp. 15–22, 2019, doi: 10.1016/j.psychres.2019.07.005.
- [65] L. V. der Maaten and G. Hinton, “Visualizing data using t-SNE.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.

- [66] C. M. Olbert, G. J. Gala, and L. A. Tupler, “Quantifying Heterogeneity Attributable to Polythetic Diagnostic Criteria: Theoretical Framework and Empirical Application,” *J. Abnorm. Psychol.*, vol. 123, no. 2, pp. 452–462, 2014, doi: 10.1037/a0036068.
- [67] E. Calli, E. Sogancioglu, E. Th. Scholten, K. Murphy, and B. van Ginneken, “Handling label noise through model confidence and uncertainty: application to chest radiograph classification,” *Méd. Imaging 2019: Comput.-Aided Diagn.*, vol. 10950, pp. 1095016-1095016–8, 2019, doi: 10.1117/12.2514290.
- [68] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, “Robust Learning at Noisy Labeled Medical Images: Applied to Skin Lesion Classification,” *2019 IEEE 16th Int. Symp. Biomed. Imaging (ISBI 2019)*, vol. 00, pp. 1280–1283, 2019, doi: 10.1109/isbi.2019.8759203.
- [69] M. Pechenizkiy, A. Tsymbal, S. Puuronen, and O. Pechenizkiy, “Class Noise and Supervised Learning in Medical Domains: The Effect of Feature Extraction,” *19th IEEE Symp. Comput.-Based Méd. Syst. (CBMS’06)*, pp. 708–713, 2006, doi: 10.1109/cbms.2006.65.
- [70] D. Gamberger, N. Lavrac, and C. Groseelj, “Experiments with Noise Filtering in a Medical Domain,” in *Proceedings of the Sixteenth International Conference on Machine Learning*, in ICML ’99. Morgan Kaufmann Publishers Inc., 1999, pp. 143–151.
- [71] S. Ji and J. Ye, “Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection,” *IEEE Trans. Neural Netw.*, vol. 19, no. 10, pp. 1768–1782, 2008, doi: 10.1109/tnn.2008.2002078.
- [72] G. B. Chand *et al.*, “Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning,” *Brain*, vol. 143, no. 3, pp. 1027–1038, 2020, doi: 10.1093/brain/awaa025.
- [73] B. Narayanan *et al.*, “Genetic Sources of Subcomponents of Event-Related Potential in the Dimension of Psychosis Analyzed From the B-SNIP Study,” *Am. J. Psychiatry*, vol. 172, no. 5, pp. 466–478, 2015, doi: 10.1176/appi.ajp.2014.13101411.
- [74] O. B. Smeland *et al.*, “Genome-wide analysis reveals extensive genetic overlap between schizophrenia, bipolar disorder, and intelligence,” *Mol Psychiatr*, vol. 25, no. 4, pp. 844–853, 2020, doi: 10.1038/s41380-018-0332-x.

- [75] J. B. Potash, “Carving chaos: genetics and the classification of mood and psychotic syndromes,” *Harvard review of psychiatry*, vol. 14, no. 2, pp. 47–63, 2006, doi: 10.1080/10673220600655780.
- [76] G. D. Pearlson, “Etiologic, Phenomenologic, and Endophenotypic Overlap of Schizophrenia and Bipolar Disorder,” *Annu Rev Clin Psycho*, vol. 11, no. 1, pp. 1–31, 2015, doi: 10.1146/annurev-clinpsy-032814-112915.
- [77] B. A. Clementz *et al.*, “Psychosis Biotypes: Replication and Validation from the B-SNIP Consortium,” *Schizophrenia Bull*, vol. 48, no. 1, pp. sbab090–, 2021, doi: 10.1093/schbul/sbab090.
- [78] W. Yin, L. Li, and F.-X. Wu, “Deep learning for brain disorder diagnosis based on fMRI images,” *Neurocomputing*, vol. 469, pp. 332–345, 2022, doi: 10.1016/j.neucom.2020.05.113.
- [79] H. Falakshahi *et al.*, “Time-varying Graphs: A Method to Identify Abnormal Integration and Disconnection in Functional Brain Connectivity with Application to Schizophrenia,” *2020 Ieee 20th Int Conf Bioinform Bioeng Bibe*, vol. 00, pp. 417–424, 2020, doi: 10.1109/bibe50027.2020.00074.
- [80] H. Falakshahi *et al.*, “Path analysis: A method to estimate altered pathways in time-varying graphs of neuroimaging data,” vol. 6, no. 3, pp. 634–664, 2022, doi: 10.1162/netn_a_00247.
- [81] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde, “Functional connectivity in the motor cortex of resting human brain using echo-planar mri,” *Magn. Reson. Med.*, vol. 34, no. 4, pp. 537–541, 1995, doi: 10.1002/mrm.1910340409.
- [82] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon, “Functional connectivity in the resting brain: A network analysis of the default mode hypothesis,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 1, pp. 253–258, 2003, doi: 10.1073/pnas.0135058100.
- [83] V. D. Calhoun and T. Adalı, “Multisubject Independent Component Analysis of fMRI: A Decade of Intrinsic Networks, Default Mode, and Neurodiagnostic Discovery,” *IEEE Rev. Biomed. Eng.*, vol. 5, pp. 60–73, 2012, doi: 10.1109/rbme.2012.2211076.

- [84] V. D. Calhoun, T. Eichele, and G. Pearlson, “Functional Brain Networks in Schizophrenia: A Review,” *Front. Hum. Neurosci.*, vol. 3, p. 17, 2009, doi: 10.3389/neuro.09.017.2009.
- [85] A. Anderson and M. S. Cohen, “Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial,” *Front. Hum. Neurosci.*, vol. 7, p. 520, 2013, doi: 10.3389/fnhum.2013.00520.
- [86] T. Kaufmann *et al.*, “Disintegration of Sensorimotor Brain Networks in Schizophrenia,” *Schizophrenia Bull*, vol. 41, no. 6, pp. 1326–1335, 2015, doi: 10.1093/schbul/sbv060.
- [87] M. R. Arbabshirani, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun, “Classification of schizophrenia patients based on resting-state functional network connectivity,” *Front. Neurosci.*, vol. 7, p. 133, 2013, doi: 10.3389/fnins.2013.00133.
- [88] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun, “Tracking Whole-Brain Connectivity Dynamics in the Resting State,” *Cereb Cortex*, vol. 24, no. 3, pp. 663–676, 2014, doi: 10.1093/cercor/bhs352.
- [89] V. D. Calhoun, R. Miller, G. Pearlson, and T. Adalı, “The Chronnectome: Time-Varying Connectivity Networks as the Next Frontier in fMRI Data Discovery,” *Neuron*, vol. 84, no. 2, pp. 262–274, 2014, doi: 10.1016/j.neuron.2014.10.015.
- [90] B. Rashid *et al.*, “Connectivity dynamics in typical development and its relationship to autistic traits and autism spectrum disorder,” *Hum. Brain Mapp.*, vol. 39, no. 8, pp. 3127–3142, 2018, doi: 10.1002/hbm.24064.
- [91] B. Rashid, E. Damaraju, G. D. Pearlson, and V. D. Calhoun, “Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects,” *Front Hum Neurosci*, vol. 8, p. 897, 2014, doi: 10.3389/fnhum.2014.00897.
- [92] B. Rashid *et al.*, “Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity,” *Neuroimage*, vol. 134, pp. 645–657, 2016, doi: 10.1016/j.neuroimage.2016.04.051.

- [93] M. Demirtaş *et al.*, “Dynamic functional connectivity reveals altered variability in functional connectivity among patients with major depressive disorder,” *Hum. Brain Mapp.*, vol. 37, no. 8, pp. 2918–2930, 2016, doi: 10.1002/hbm.23215.
- [94] H. Rokham, G. Pearlson, A. Abrol, H. Falakshahi, S. Plis, and V. D. Calhoun, “Addressing Inaccurate Nosology in Mental Health: A Multilabel Data Cleansing Approach for Detecting Label Noise From Structural Magnetic Resonance Imaging Data in Mood and Psychosis Disorders,” *Biological Psychiatry Cognitive Neurosci Neuroimaging*, vol. 5, no. 8, pp. 819–832, 2020, doi: 10.1016/j.bpsc.2020.05.008.
- [95] Y. Du *et al.*, “NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders,” *Neuroimage Clin*, vol. 28, p. 102375, 2020, doi: 10.1016/j.nicl.2020.102375.
- [96] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008, doi: 10.1093/biostatistics/kxm045.
- [97] Y. Tu *et al.*, “Abnormal thalamocortical network dynamics in migraine,” *Neurology*, vol. 92, no. 23, pp. e2706–e2716, 2019, doi: 10.1212/wnl.0000000000007607.
- [98] Z. Fu *et al.*, “Characterizing dynamic amplitude of low-frequency fluctuation and its relationship with dynamic functional connectivity: An application to schizophrenia,” *NeuroImage*, vol. 180, no. Pt B, pp. 619–631, 2018, doi: 10.1016/j.neuroimage.2017.09.035.
- [99] E. Fiorenzato *et al.*, “Dynamic functional connectivity changes associated with dementia in Parkinson’s disease,” *Brain*, vol. 142, no. 9, pp. 2860–2872, 2019, doi: 10.1093/brain/awz192.
- [100] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [101] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [102] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.

- [103] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *Arxiv*, 2014, doi: 10.48550/arxiv.1412.6980.
- [104] G. C. Cawley and N. L. C. Talbot, “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation,” *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.
- [105] G. Collin, H. E. H. Pol, S. V. Haijma, W. Cahn, R. S. Kahn, and M. P. van den Heuvel, “Impaired Cerebellar Functional Connectivity in Schizophrenia Patients and Their Healthy Siblings,” *Front. Psychiatry*, vol. 2, p. 73, 2011, doi: 10.3389/fpsyt.2011.00073.
- [106] L. Wang *et al.*, “Disruptive changes of cerebellar functional connectivity with the default mode network in schizophrenia,” *Schizophr. Res.*, vol. 160, no. 1–3, pp. 67–72, 2014, doi: 10.1016/j.schres.2014.09.034.
- [107] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” *Inform Fusion*, vol. 81, pp. 84–90, 2022, doi: 10.1016/j.inffus.2021.11.011.
- [108] R. de Filippis *et al.*, “Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: a systematic review,” *Neuropsychiatr. Dis. Treat.*, vol. 15, pp. 1605–1627, 2019, doi: 10.2147/ndt.s202418.
- [109] M. Picó-Pérez, R. Vieira, M. Fernández-Rodríguez, M. A. P. D. Barros, J. Radua, and P. Morgado, “Multimodal meta-analysis of structural gray matter, neurocognitive and social cognitive fMRI findings in schizophrenia patients,” *Psychol. Med.*, vol. 52, no. 4, pp. 614–624, 2022, doi: 10.1017/s0033291721005523.
- [110] M. P. van den Heuvel and A. Fornito, “Brain Networks in Schizophrenia,” *Neuropsychol Rev*, vol. 24, no. 1, pp. 32–48, 2014, doi: 10.1007/s11065-014-9248-7.
- [111] K. J. Friston *et al.*, “Analysis of fMRI Time-Series Revisited,” *Neuroimage*, vol. 2, no. 1, pp. 45–53, 1995, doi: 10.1006/nimg.1995.1007.
- [112] M. Symms, H. R. Jäger, K. Schmierer, and T. A. Yousry, “A review of structural magnetic resonance neuroimaging,” *J Neurology Neurosurg Psychiatry*, vol. 75, no. 9, p. 1235, 2004, doi: 10.1136/jnnp.2003.032714.

- [113] M. Liang *et al.*, “Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging,” *NeuroReport*, vol. 17, no. 2, p. 209, 2006, doi: 10.1097/01.wnr.0000198434.06518.b8.
- [114] S. Goswami *et al.*, “A preliminary study to investigate resting state fMRI as a potential group differentiator for schizophrenia,” *Asian J. Psychiatry*, vol. 52, p. 102095, 2020, doi: 10.1016/j.ajp.2020.102095.
- [115] Y. Zhou *et al.*, “Functional disintegration in paranoid schizophrenia using resting-state fMRI,” *Schizophr. Res.*, vol. 97, no. 1–3, pp. 194–205, 2007, doi: 10.1016/j.schres.2007.05.029.
- [116] D. B. Keator *et al.*, “The Function Biomedical Informatics Research Network Data Repository,” *Neuroimage*, vol. 124, no. Pt B, pp. 1074–1079, 2016, doi: 10.1016/j.neuroimage.2015.09.003.
- [117] Y.-D. Zhang *et al.*, “Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation,” *Inf. Fusion*, vol. 64, pp. 149–187, 2020, doi: 10.1016/j.inffus.2020.07.006.
- [118] J. Sui, R. Jiang, J. Bustillo, and V. Calhoun, “Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises,” *Biol Psychiat*, vol. 88, no. 11, pp. 818–828, 2020, doi: 10.1016/j.biopsych.2020.02.016.
- [119] Y. Du, Z. Fu, and V. D. Calhoun, “Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging,” *Front. Neurosci.*, vol. 12, p. 525, 2018, doi: 10.3389/fnins.2018.00525.
- [120] D. Bzdok and A. Meyer-Lindenberg, “Machine Learning for Precision Psychiatry: Opportunities and Challenges,” *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018, doi: 10.1016/j.bpsc.2017.11.007.
- [121] L. A. Clark, D. Watson, and S. Reynolds, “Diagnosis and Classification of Psychopathology: Challenges to the Current System and Future Directions,” *Annu. Rev. Psychol.*, vol. 46, no. 1, pp. 121–153, 1995, doi: 10.1146/annurev.ps.46.020195.001005.

- [122] M. L. Phillips and D. J. Kupfer, “Bipolar disorder diagnosis: challenges and future directions,” *Lancet*, vol. 381, no. 9878, pp. 1663–1671, 2013, doi: 10.1016/s0140-6736(13)60989-7.
- [123] M. C. O’Donovan and M. J. Owen, “The implications of the shared genetics of psychiatric disorders,” *Nat. Med.*, vol. 22, no. 11, pp. 1214–1219, 2016, doi: 10.1038/nm.4196.
- [124] P. Cuijpers, “Targets and outcomes of psychotherapies for mental disorders: an overview,” *World Psychiatry*, vol. 18, no. 3, pp. 276–285, 2019, doi: 10.1002/wps.20661.
- [125] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, “Learning From Noisy Labels With Deep Neural Networks: A Survey,” *Ieee T Neur Net Lear*, vol. PP, no. 99, pp. 1–19, 2022, doi: 10.1109/tnnls.2022.3152527.
- [126] A. Jablensky, “Psychiatric classifications: validity and utility,” *World Psychiatry*, vol. 15, no. 1, pp. 26–31, 2016, doi: 10.1002/wps.20284.
- [127] C. C. Conway *et al.*, “Rethinking the Diagnosis of Mental Disorders: Data-Driven Psychological Dimensions, Not Categories, as a Framework for Mental-Health Research, Treatment, and Training,” *Curr Dir Psychol Sci*, vol. 30, no. 2, pp. 151–158, 2021, doi: 10.1177/0963721421990353.
- [128] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/bf00058655.