ADVANCED BOUNDARY-ENHANCED INSTANCE SEGMENTATION AND SPATIAL-TEMPORAL TRANSFORMER MODELS FOR AUTOMATED SCHIZOPHRENIC INVESTIGATION

OSASUMWEN RAPHAEL IMARHIAGBE
Bachelor of Science, Computer Science.
Benson Idahosa University, 2022

A thesis submitted
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

Department of Mathematics and Computer Science
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

ADVANCED BOUNDARY-ENHANCED INSTANCE SEGMENTATION AND SPATIAL-TEMPORAL TRANSFORMER MODELS FOR AUTOMATED SCHIZOPHRENIC INVESTIGATION

OSASUMWEN RAPHAEL IMARHIAGBE

Date of Defence: August 12, 2025

Dr. John Z. Zhang Thesis Supervisor	Associate Professor	Ph.D.
Dr. Wendy Osborn Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. Yllias Chali Thesis Examination Committee Member	Professor	Ph.D.
Dr. Andrew Fiori Chair, Thesis Examination Committee	Associate Professor	Ph.D.

Dedication

To my mother and siblings, for their endless love and encouragement, and to all those affected by schizophrenia, whose strength and resilience inspired this research

Abstract

Accurate segmentation and detection in neuroimaging is essential for advancing clinical understanding and the diagnosis of schizophrenia. This thesis introduces Boundary-Refined Attention Network (BoRefAttnNet), a novel boundary-refined 3D U-Net variant specifically designed for precise segmentation of subcortical brain structures from structural magnetic resonance imaging (sMRI). BoRefAttnNet incorporates multi-scale boundary attention modules that explicitly highlight anatomically critical edges while suppressing background noise, significantly improving segmentation accuracy for small or complex anatomical structures. Evaluations using FastSurfer-processed sMRI data from the publicly available Centre for Biomedical Research Excellence (COBRE) dataset demonstrate that BoRefAttnNet significantly outperforms conventional 3D U-Net baselines in accurately delineating key subcortical structures, including the hippocampus, amygdala, and basal ganglia.

Building upon this enhanced segmentation capability, we further experiment with a Dynamic Spatial-Temporal Transformer Model (DySTTM) to detect schizophrenia by integrating structural and functional MRI (fMRI) modalities. The DySTTM leverages spatial attention to capture anatomical interdependencies from segmented sMRI data and temporal attention to model dynamic brain connectivity patterns from resting-state fMRI. Experimental results indicate that the integration of these multimodal imaging features using DySTTM provides superior diagnostic accuracy and interpretability compared to established models such as 3D ResNet and XGBoost classifiers.

Acknowledgments

I extend my profound gratitude to my supervisor, Dr. John Z. Zhang, for his invaluable guidance, continuous support, and insightful advice throughout my research. His continued guidance significantly shaped my research trajectory.

I sincerely appreciate members of my thesis examination committee, Dr. Wendy Osborn and Dr. Yllias Chali, for their thoughtful feedback. Special thanks to Dr. Andrew Fiori for chairing the examination committee and facilitating a smooth examination process.

I am grateful to Alberta Innovates, the University of Lethbridge School of Graduate Studies, Alberta Machine Intelligence Institute, and the Digital Research Alliance of Canada for the respective grants and awards which significantly aided my research efforts.

I acknowledge the impactful courses offered by Dr. Daya Gaur, Dr. Robert Benkoczi, and Dr. John Zhang, which provided crucial foundations for my research. My experience as a Teaching Assistant under Dr. Daya Gaur, Dr. John Sheriff, Dr. John Zhang, Milad Fakhari, and Jannatul Maowa was enriching as well, and I thank them for their invaluable guidance.

I deeply appreciate my colleagues, Olanrewaju Oladokun, Asif Talukdar, Naveen Kumar Vadlamudi, Reza Ardestani, Al Hasib Mahamud, Juhyoung Park, Md. Moshin Uddin, and Sudip Pokhrel, for engaging and insightful discussions on research, emerging technologies, and trends which in turn greatly enriched my understanding and perspectives.

Lastly, I thank my family deeply for their unwavering love, support, and encouragement throughout my academic journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Recent advances in computational methodologies, particularly deep learning, computer vision, and medical imaging, have significantly enhanced diagnostic accuracy and clinical decision-making processes. Deep learning, a subset of machine learning inspired by neural structures in biological brains, leverages hierarchical artificial neural networks to automatically extract intricate features from complex data [1]. Techniques such as convolutional neural networks (CNNs) have excelled in tasks involving image and video data, notably object recognition and medical image segmentation [3, 60]. Graph neural networks (GNNs) have also emerged, demonstrating remarkable capabilities in capturing relational and structural information inherent in brain connectivity and functional imaging data [14]. Transformer architectures, initially introduced for natural language processing, have further revolutionized deep learning by effectively modeling long-range dependencies and contextual interactions, making them highly suitable for complex tasks such as medical image analysis and disease detection [12, 14].

Despite these advancements, considerable challenges remain unresolved within medical imaging, particularly concerning brain Magnetic Resonance Imaging (MRI). Brain MRI data poses unique computational hurdles, including high dimensionality, significant inter-subject variability, subtle pathological features, and typically low signal-to-noise ratios [3, 34]. These complexities necessitate sophisticated machine learning frameworks capable of extracting nuanced, clinically relevant insights while ensuring robustness and

generalizability across diverse patient populations [31, 8].

Schizophrenia, a complex psychiatric disorder characterized by subtle yet distinctive structural and functional brain abnormalities, exemplifies these computational challenges [2, 8]. Accurate automated detection and diagnosis require advanced computational methods capable of integrating multimodal imaging data—combining structural MRI (sMRI) and functional MRI (fMRI)—to capture comprehensive insights into schizophrenia pathology [8]. This thesis introduces computational methods, specifically Transformer-based architectures and attention mechanisms, explicitly tailored to the spatial-temporal integration of multimodal MRI data. These methods significantly enhance diagnostic accuracy, robustness, and interpretability, addressing critical gaps left by conventional methodologies [12, 14].

1.2 Problem Statement

Automated computational detection of schizophrenia is hindered by:

- **Segmentation Precision:** Current methodologies struggle to accurately segment brain structures due to anatomical variability and complex spatial boundaries inherent in MRI data.
- **Multimodal Integration Complexity:** Existing methods inadequately integrate structural MRI with dynamic functional MRI, limiting their ability to capture comprehensive biomarkers.

This thesis addresses these challenges by developing and validating sophisticated, interpretable computational models specifically optimized for multimodal schizophrenia detection.

1.3 Research Objectives and Contributions

This thesis substantially contributes to computational medical imaging and schizophrenia research by introducing an innovative Segmentation-based computational model for precise brain Magnetic Resonance Imaging segmentation, experimenting on a robust framework for the integration of structural and functional MRI using advanced attention mechanisms with the aim of demonstrating superior detection performance and robustness compared to existing computational techniques.

Further, this thesis shows promise in enriching the computational tools available to clinicians, significantly advancing diagnostic precision and interpretability in schizophrenia research. Succinctly, the objectives are:

1. Develop precise boundary-aware segmentation methods for anatomical brain regions using structural MRI.

2. Experiment using Transformer-based computational models capable of integrating spatial-temporal features derived from structural and functional MRI.

3. Validate these methodologies rigorously against current state-of-the-art computational benchmarks.

1.4 Overview of Methodology

To address these objectives, this research employs a structured computational approach encompassing several interconnected stages:

- **Data Preparation and Preprocessing:** Preprocessing methods include skull stripping, normalization, artifact correction, registration, and alignment of multimodal datasets for consistency.

- **Advanced Segmentation Techniques:** A novel Boundary-refined attention segmentation network, BoRefAttnNet, was developed to accurately delineate anatomical boundaries within structural MRI.

- **Dynamic Spatial-Temporal Transformer Model:** A dual-stream Transformer architecture that integrates segmented structural MRI features with functional MRI time-series data was employed, capturing both spatial and temporal schizophrenia biomarkers.

1.5 Thesis Outline

This thesis is organized into seven chapters. Chapter 1 introduces the research problem, outlining the motivation, objectives, and overall contributions of this work.

Chapter 2 provides the necessary background, presenting key concepts in schizophrenia, neuroimaging techniques, and computational methods such as neural networks, deep learning, and transformers. Chapter 3 then surveys related works, offering a comprehensive review of existing literature on computational methods in schizophrenia neuroimaging and identifying critical research gaps.

Chapter 4 details our first methodological contribution, the BoRefAttnNet framework, a boundary-refined segmentation approach designed to improve precision in delineating complex brain structures. This chapter also includes comparative experiments against benchmark architectures.

In Chapter 5, we present the Dynamic Spatial-Temporal Transformer Model, which integrates multimodal MRI data for schizophrenia investigation. We demonstrate its effectiveness through experiments, reporting both quantitative results and statistical validation.

Finally, Chapter 6 discusses the implications of our findings, summarizes the key contributions of the thesis, and outlines promising directions for future research.

Chapter 2

Background

This chapter provides a structured overview emphasizing the role of computational artificial intelligence (AI), particularly deep learning, in enhancing schizophrenia research through medical imaging. We discuss schizophrenia, explain relevant medical imaging techniques, and review essential computational approaches, explicitly focusing on deep learning and transformers. We conclude by summarizing existing approaches and positioning this research within the broader landscape of computational methods in schizophrenic investigation.

2.1 Computational Artificial Intelligence

2.1.1 Neural Networks

Artificial neural networks (ANNs) are computational models inspired by biological neural systems found in living organisms. As outlined by Aggarwal [1], these networks simulate biological mechanisms by employing interconnected computational units referred to as neurons. Unlike biological neurons connected via axons and dendrites through synapses, artificial neurons connect through weighted links that mimic synaptic strengths. In artificial neural networks, each computational neuron as illustrated in Figure 2.1, receives inputs scaled by weights, influencing the neuron's output through mathematical operations. The computed values propagate from input neurons through hidden layers, eventually reaching output neurons, allowing the network to produce predictions or classifications based on input data.

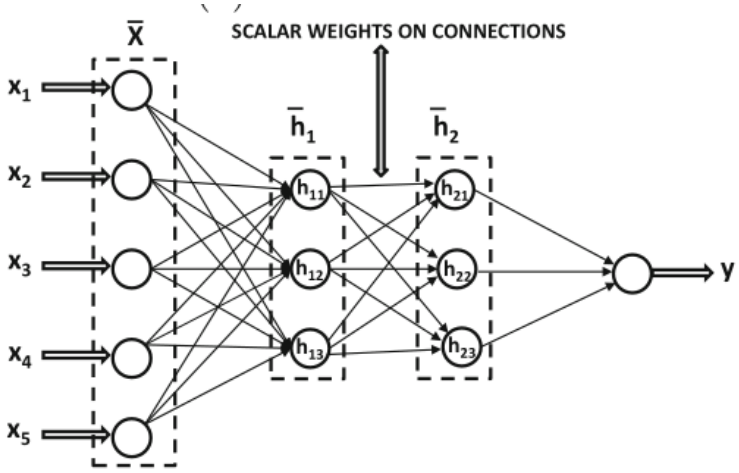


Figure 2.1: Typical architecture of a multi-layer neural network illustrating scalar weights on connections between input, hidden, and output layers (Adapted from Aggarwal [1]).

Learning Mechanism

Learning in neural networks involves adjusting connection weights based on input-output training examples. Each training instance contains input data (features) and corresponding desired outputs (labels). During training, prediction errors (differences between actual and predicted outputs) guide the iterative refinement of weights to minimize these errors. This training paradigm enables neural networks to achieve high performance in diverse tasks such as classification, regression, and complex pattern recognition [1].

The effectiveness of neural networks primarily lies in their ability to generalize, accurately predicting outcomes for unseen data, having only learned from defined training data. This property, known as *model generalization*, is a cornerstone of their broad applicability and practical utility.

Aggarwal [1] emphasizes that neural networks can be viewed as *computational graphs*, where nodes represent neurons or computation units, and edges represent weighted connections. This graph structure facilitates understanding, visualizing, and implementing neural network computations systematically. Each node performs simple operations, and the overall complexity and learning capacity emerge from the structured interactions of these elementary units.

Mathematical Formulation

According to Aggarwal [1], neural networks aim to learn a predictive function $f(\cdot)$ from input data $\overline{X}$ to outputs y:

$$y = f_{\overline{W}}(\overline{X}) \tag{2.1}$$

where $\overline{W}$ represents the learned weights of the network. The training process involves optimizing the weights to minimize a defined loss function, typically represented as:

$$\min_{\overline{W}} \text{loss}(y, f_{\overline{W}}(\overline{X})) \tag{2.2}$$

Activation Functions

Activation functions introduce non-linearities into neural networks, enabling them to model complex relationships. Common activation functions include Sigmoid [61], Tanh [43], and ReLU [49, 70]

- $\sigma(v) = \frac{1}{1+e^{-v}}$, derivative: $\sigma(v)(1-\sigma(v))$
- $\tanh(v) = \frac{e^{2v}-1}{e^{2v}+1}$, derivative: $1-\tanh^2(v)$
- $\text{ReLU}(v) = \max(v, 0)$, derivative: 1 for $v > 0$, else 0

These functions enable neural networks to learn more sophisticated representations and decision boundaries than linear models [1]. Also, common loss functions such as mean squared error for regression problems and cross-entropy loss for classification tasks, guide learning towards optimal predictive accuracy.

2.1.2 Deep Learning

Deep learning, a sophisticated branch of machine learning, extends classical neural networks by introducing deeper, more complex hierarchical architectures that excel at automated feature extraction and representation learning.

Deep neural networks (DNNs), by virtue of their multiple layers, have the capacity to learn increasingly abstract representations of data. This property is most evident in structured data domains such as images and language, where architectures like Convolutional Neural Networks (CNNs) or Transformers explicitly exploit hierarchical feature learning. However, in general DNNs, the degree of hierarchy in learned features may vary depending on architecture and application. Each subsequent layer extracts increasingly abstract features, significantly enhancing the network's capacity to capture intricate data structures [1].

A Convolutional neural network (CNN), detailed by Aggarwal [1] in Figure 2.2, exemplifies such architecture, it illustrates the basic structure of one of the earliest convolutional neural networks. The network begins with a grayscale input image of size 32×32 pixels which passes through the first convolutional layer (C1), and then applies six 5×5 filters to produce six 28×28 feature maps. A subsampling (S2) or average pooling layer then reduces each feature map to 14×14, stabilizing feature extraction while lowering computational cost. The next convolutional layer (C3) increases the number of feature maps to 16, followed again by a subsampling layer (S4) that reduces the resolution to 5×5. At this stage, the network transitions into a fully connected structure: layer C5 (with 120 units), followed by layer F6 (84 units), and finally an output layer (O) with 10 neurons corresponding to class predictions.

Notably, subsampling layers in the original LeNet-5 used average pooling with trainable weights and biases — a design choice that differs from modern CNNs, which typically use max-pooling. Activation functions (sigmoid in the original) were applied after each convolution and subsampling stage but are not shown explicitly in Aggarwal's diagram.

Patch-Based Training

Patch-based training has become indispensable in medical image segmentation tasks due to limited GPU memory and the necessity to learn detailed local features. Instead

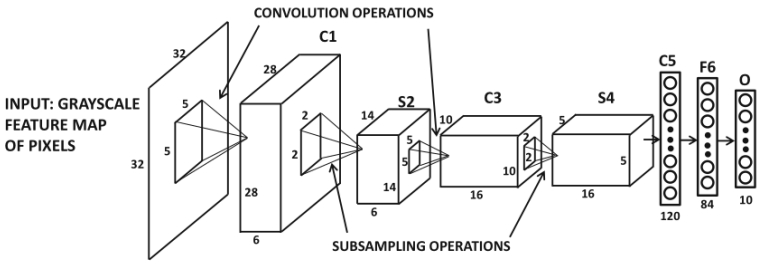


Figure 2.2: Detail architectural representation of a convolutional neural network illustrating subsampling layers (Adapted from Aggarwal [1]).

of processing entire images simultaneously, this method addresses the GPU memory constraints of 3D medical image segmentation by dividing volumetric scans into smaller subvolumes (patches). These patches are individually processed by the network, enabling detailed learning of local features while keeping memory usage feasible. After inference, the predictions from patches are reassembled to reconstruct the segmentation of the full image. Proper choice of patch size and sampling strategy is essential to preserve spatial context and ensure accurate delineation of anatomical structures.

This approach was notably employed in the MICCAI 2013 Grand Challenge for brain tumor segmentation, demonstrating superior performance compared to whole-image methodologies [44, 15]. However, careful optimization of patch size and sampling strategy is required to preserve contextual information critical for accurate segmentation.

Batch Normalization

Batch normalization stabilizes and accelerates the training of deep neural networks (DNNs) by normalizing inputs to layers, significantly reducing internal covariate shift. Given a layer's inputs x_i, batch normalization computes normalized outputs y_i as follows:

$$y_i = \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} + \beta$$

where, μ_B and σ_B^2 are the mean and variance of the mini-batch, respectively, while γ and β are learnable parameters enabling the model to adaptively scale and shift normalized

outputs [35, 4]. Batch normalization has been shown to effectively mitigate the vanishing gradient problem, which arises during backpropagation in deep networks when gradients are repeatedly multiplied by small values across many layers (particularly with sigmoid or tanh activations). As a result, the gradient can shrink to near zero in earlier layers, preventing their weights from being updated and hindering learning. By stabilizing these gradients, batch normalization facilitates faster convergence and improves the generalization ability of deep models [1].

Adam Optimizer

The Adaptive Moment Estimation (Adam) optimizer [40] is particularly effective in handling sparse gradients, a situation where most gradient values are zero (or near zero) and only a few are non-zero. Such sparsity commonly arises in models with sparse inputs such as one-hot encoded text or when regularization methods like L1 encourage sparsity. In these cases, only a small subset of weights receive updates during training. Adam addresses this challenge by adaptively scaling the learning rate for each parameter, using estimates of both the first and second moments of the gradients. This ensures that even sparse and irregular updates are applied effectively, leading to more stable and efficient training. The weight updates for Adam are expressed as:

$$w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates of the first and second gradient moments, respectively [1]. While Adam has shown substantial benefits in accelerating convergence, its standard form can sometimes generalize poorly compared to classical stochastic gradient descent. Improved variants, such as Normalized Direction-preserving Adam (ND-Adam), address these issues, enhancing both convergence and generalization [76, 5].

ReLU Activation

The Rectified Linear Unit (ReLU) activation function, defined as:

$$\text{ReLU}(x) = \max(0, x)$$

has gained popularity due to its computational simplicity and effectiveness in avoiding vanishing gradient issues common with sigmoid and hyperbolic tangent functions. It introduces sparsity by activating neurons selectively, enhancing training efficiency and convergence [49, 70]. Despite these advantages, standard ReLU activations may suffer from dying ReLU issues, where neurons permanently deactivate. Variants such as Leaky ReLU [20] and Parametric ReLU (PReLU) [30] mitigate these concerns by maintaining neuron activity for negative inputs.

2.1.3 Transformer Models

Transformer models have emerged as a powerful architecture in machine learning, initially proposed to address limitations in capturing long-range dependencies inherent in sequence data. Originally introduced by Vaswani et al. [68] for Natural Language Processing (NLP), Transformers have since been adapted and extensively utilized in Computer Vision, significantly enhancing the ability to model complex spatial relationships in images and videos [1]. The key innovation of Transformer models is the self-attention mechanism, which allows the model to dynamically weigh the importance of input data points relative to each other. Unlike traditional Convolutional Neural Networks (CNNs), which apply localized receptive fields, Transformers can model global dependencies across the entire input space.

Transformers in Computer Vision

Transformers have been successfully adapted from NLP to computer vision tasks through models like Vision Transformer (ViT), introduced by Dosovitskiy et al. [19]. Instead of pro-

cessing pixels directly, ViT divides images into fixed-size patches, embedding these patches into vectors processed by Transformer blocks. This enables the model to capture complex spatial relations and global context within visual data, overcoming limitations inherent to CNN architectures such as restricted receptive fields and limited global context awareness [1].

2.2 Clinical and Imaging Context for Deep Learning in Schizophrenia

2.2.1 Schizophrenia

Schizophrenia is a chronic and severe psychiatric disorder characterized by fundamental distortions in thinking, perception, emotion, and behavior [58]. Typically emerging in late adolescence or early adulthood, schizophrenia features episodes of psychosis—manifested by delusions and hallucinations—alongside persistent impairments in social and occupational functioning.

Clinically, schizophrenic symptoms are grouped into three domains: positive symptoms (such as delusions, hallucinations, disorganized speech, and behavior), negative symptoms (including diminished emotional expression, alogia, avolition, anhedonia, and social withdrawal), and cognitive symptoms (notably deficits in attention, working memory, executive functioning, and processing speed). These cognitive impairments often precede psychotic episodes and significantly influence long-term functional outcomes, highlighting the complexity and pervasive impact of the disorder [58, 45].

2.2.2 Basic Concepts in Medical Imaging

Medical imaging, specifically Magnetic Resonance Imaging (MRI), provides essential data used to analyze anatomical and functional properties of the brain, serving as critical inputs for computational analyses, particularly in the context of schizophrenic research. Understanding fundamental concepts of MRI image acquisition, preprocessing, and segmentation techniques is crucial for effectively utilizing these images in machine learning

and deep learning workflows.

MRI is a non-invasive imaging modality widely utilized in neuroscience research due to its superior soft tissue contrast, high spatial resolution, and functional imaging capabilities. MRI generates images based on the nuclear magnetic resonance (NMR) properties of hydrogen nuclei in biological tissues, primarily influenced by relaxation times (T1, T2) and proton density.

MRI involves placing the subject within a strong magnetic field, causing hydrogen protons in the body's water molecules to align with this field. Radiofrequency (RF) pulses disturb this alignment, and as protons return to equilibrium, they emit signals detected by MRI scanners. Differences in proton relaxation times and densities among various tissues produce distinct image contrasts. Two primary relaxation processes, T1 (longitudinal relaxation) and T2 (transverse relaxation), form the basis of different MRI image contrasts.

Pulse Sequences: T1-weighted, T2-weighted, FLAIR

MRI pulse sequences are used to highlight different tissue characteristics:

- **T1-weighted images (T1w)**: Provide clear anatomical details by emphasizing differences in longitudinal relaxation times (the time it takes protons to realign with the main magnetic field after excitation). T1w images are extensively used to study brain morphology, structural abnormalities, and tissue volumes in schizophrenia analysis [69].

- **T2-weighted images (T2w)**: Emphasize differences in transverse relaxation times (the time it takes protons to lose phase coherence in the transverse plane, leading to signal decay), highlighting pathological tissues with increased water content. T2w images are commonly used for clinical assessment of lesions and tissue abnormalities.

- **FLAIR (Fluid-Attenuated Inversion Recovery)**: Suppresses cerebrospinal fluid (CSF) signals, enhancing visualization of pathological lesions such as demyelina-

tion and subtle cortical abnormalities, which are useful in the analysis of psychiatric conditions and their neuroanatomical correlates.

Spatial Resolution and Voxel Considerations

MRI images are represented in three-dimensional grids of volume elements known as voxels. Spatial resolution, defined by voxel size, critically affects the accuracy and detail of MRI analysis. Higher resolution images (smaller voxels) provide greater anatomical detail but increase acquisition time and computational resources required for processing. Conversely, lower resolution (larger voxels) images reduce computational demands and imaging time but may miss subtle structural details important in identifying fine-grained pathological changes, particularly relevant in schizophrenia imaging studies. Thus, choosing appropriate voxel dimensions is essential to balance computational feasibility with clinical diagnostic requirements [3].

2.2.3 Neuroimaging Techniques

Neuroimaging encompasses multiple modalities that provide complementary views of brain structure and function, which are crucial for computational modeling of schizophrenia. The two most relevant modalities discussed in our research are structural Magnetic Resonance Imaging (sMRI) and functional Magnetic Resonance Imaging (fMRI). These techniques produce high-dimensional representations of the brain, capturing its complex anatomy and activity patterns [14]. In machine learning (ML) and deep learning workflows, data from these modalities serve as rich input features, each with distinct spatial, temporal, or spectral characteristics that can be leveraged for tasks such as segmentation and classification. The Scope of this research focuses on structural MRI (sMRI) and functional MRI (fMRI)

Structural MRI (sMRI). Structural MRI (typically T1-weighted scans) provides high-resolution 3D images of brain anatomy, depicting the volumetric distribution of gray matter, white matter, and cerebrospinal fluid. Each sMRI scan is a volumetric intensity map (with

$\sim$1 mm isotropic voxels) capturing the spatial structure of the brain at a single time point (i.e., no temporal component). These volumetric images are commonly used as input to convolutional neural networks (CNNs) for tasks like automated brain tissue segmentation and lesion detection, as well as disease classification. For example, 3D CNN architectures have been trained on whole-brain T1-weighted MRI volumes to distinguish schizophrenia patients from healthy controls [75].

Such models can learn subtle morphological alterations (e.g. regional volume or shape differences) associated with schizophrenia. In segmentation contexts, deep networks (including U-Net variants and Transformers) can delineate structures like hippocampi or detect tumors within sMRI scans, often outperforming traditional methods by capturing global context and long-range dependencies in the 3D data [14] .

Overall, sMRI provides detailed spatial features (voxel intensities or region volumes) that feed into ML models, usually via 2D or 3D CNN pipelines, to yield predictions or anatomical labels.

Functional MRI (fMRI). Functional MRI (usually BOLD fMRI) measures brain activity over time by capturing changes in blood oxygenation, producing a 4D dataset (3D brain volume across a temporal sequence). An fMRI scan offers moderate spatial resolution (on the order of 2–3 mm voxels covering the whole brain) and an important temporal dimension (time resolution 1–2 seconds, over many minutes of scanning). The raw fMRI output is a time series of volumetric images, which can be processed into features like voxel-wise time series or region-wise averaged signals.

A common approach is to parcellate the brain into regions of interest (ROIs) and compute functional connectivity matrices, where each matrix element represents the statistical correlation between the fMRI time series of two regions. This yields a functional network per subject, which can serve as input to graph-based learning methods. For instance, one can represent fMRI data as a graph with nodes as ROIs and edges weighted by their pairwise correlations, effectively constructing a functional connectivity network [57]. Such

representations enable the use of graph neural networks (GNNs) or graph Transformers to learn from brain connectivity patterns.

Moreover, spatio-temporal deep learning models (e.g. CNN-LSTM hybrids or transformer models) can directly ingest the fMRI time-series data to capture dynamic brain activity sequences. These techniques have been applied to classify mental disorders from fMRI data by learning disease-related connectivity disturbances or activation patterns. In schizophrenia research, resting-state fMRI inputs are often used to train classifiers that differentiate patients from controls based on connectivity abnormalities, sometimes achieving promising accuracy when complex temporal dependencies are modeled.

In summary, fMRI provides time-series data (voxel or ROI-level) that can be transformed into features like time-series signals or connectivity matrices for input into temporal CNNs, RNNs, or GNNs aimed at brain-state classification and prediction tasks.

2.2.4 Image Preprocessing

Image preprocessing is an essential stage to ensure data quality, consistency, and reliability for subsequent computational analyses. This typically involves skull stripping, spatial normalization, and intensity correction procedures.

Skull Stripping

Skull stripping removes non-brain tissues such as skull, scalp, and meninges from MRI images, which is essential for isolating brain structures for subsequent computational tasks. Techniques such as AFNI's sswarper provide automated, robust, and reliable methods to segment the brain from MRI scans, enhancing accuracy and consistency for subsequent segmentation and registration steps[18].

Normalization and Registration to Standard Space

Normalization and registration techniques transform individual MRI images into a common reference framework known as the Montreal Neurological Institute (MNI) space. The

MNI space is a standardized brain template created from averaging MRI scans of many individuals, providing a universal coordinate system for brain anatomy. This standardization ensures that anatomical structures from different subjects are aligned to the same spatial reference, enabling consistent localization and comparison of brain regions across studies.

By reducing anatomical variability, MNI normalization facilitates group-level analysis, supports reproducibility, and improves the performance of computational models. For machine learning and deep learning applications, this alignment is especially critical because it provides uniform input dimensions and anatomical correspondences, ensuring accurate predictions and meaningful cross-subject comparisons [73].

Intensity Normalization and Bias Correction

Intensity normalization adjusts MRI images to achieve consistent intensity scales, reducing scanner-induced variability. Bias field correction methods, such as those provided by tools like FastSurfer, correct low-frequency intensity non-uniformities caused by magnetic field variations, significantly improving image quality and reliability for segmentation algorithms[31]

Preprocessing and Common Challenges. All neuroimaging modalities require extensive preprocessing and standardization before they can be used in learning algorithms. Raw MRI data often contain noise, artifacts (e.g., head motion, physiological noise), and site-specific variations that must be corrected. Typical preprocessing steps include slice timing and motion correction for fMRI; eddy-current and motion correction for DWI; spatial smoothing and filtering; skull stripping; and bias field correction for structural MRI. Images are usually spatially normalized to a standard template space (such as the MNI atlas) so that corresponding voxels or regions align across individuals. This reduces inter-subject variability and ensures that learned features correspond to the same anatomical loci across subjects [34]. Intensity normalization and feature scaling are also applied so that input data have consistent ranges.

Even after preprocessing, neuroimaging data remain extremely high-dimensional (e.g., tens of thousands of voxels per scan or large connectivity matrices), which poses challenges for machine learning. A major issue is the curse of dimensionality combined with limited sample sizes — clinical neuroimaging datasets for schizophrenia typically range from dozens to a few hundred subjects, which is relatively small for training deep networks. This imbalance can lead to overfitting if not addressed. Researchers mitigate this by using data augmentation, transfer learning from large neuroimaging databases, or dimensionality reduction techniques to make learning tractable [63].

Another challenge is the heterogeneity and variability in scans: differences in MRI scanners or protocols across sites can introduce unwanted variation, so harmonization methods or site-invariant models are often required. Moreover, the subtlety of disease-related patterns (signal-to-noise issues) means models must be robust to noise and confounds. Despite these obstacles, careful preprocessing and the development of specialized architectures (e.g., 3D CNNs, spatio-temporal transformers, and GNNs that respect the data's structure) have enabled increasingly successful applications of machine learning to neuroimaging. Nonetheless, ensuring robust generalization and clinically meaningful interpretation of models remains an active area of research, as high-dimensional brain data and limited samples demand diligent validation and model interpretability tools to confirm that learned features align with known neuroscience [75].

Chapter 3

Related Works

The analysis of schizophrenia through neuroimaging has experienced significant advancement over recent years. Computational techniques have evolved from traditional statistical and manual methods toward advanced machine learning and deep learning algorithms. Early studies primarily utilized structural imaging modalities and basic analytical frameworks. However, recent research trends increasingly focus on multimodal integration, advanced segmentation approaches, deep learning methods, boundary-aware techniques, and dynamic spatial-temporal analyses to improve diagnostic accuracy and clinical applicability [14]. Research trends indicate a pronounced shift towards incorporating deep learning frameworks, particularly convolutional neural networks (CNNs), graph neural networks (GNNs), and transformer architectures, for automated feature extraction and complex data modeling. These methods have significantly improved precision and interpretability in schizophrenia analysis, reflecting broader technological and methodological advances within computational neuroscience.

3.1 Traditional Computational Segmentation Techniques

Traditional computational methods for image segmentation predominantly utilize intensity and spatial features within medical images. Common techniques include thresholding (particularly Otsu's method), region growing, active contour models, the random walker algorithm, and mean shift clustering. Each of these methods leverages different assumptions about image properties and exhibits unique advantages and limitations.

3.1.1 Thresholding (Otsu's Method)

Thresholding is one of the simplest and most widely used techniques in medical imaging segmentation. It partitions an image into regions based on pixel intensity values. Otsu's method, in particular, is a popular thresholding technique that determines an optimal threshold value by maximizing inter-class variance between pixel intensities, effectively separating foreground and background components [53]. Otsu's thresholding is computationally efficient and requires minimal parameter tuning, making it suitable for scenarios where clear intensity differences exist between targeted regions, such as differentiating brain tissues (grey matter, white matter, cerebrospinal fluid). However, its primary limitation is sensitivity to intensity variations and noise, often necessitating pre-processing steps such as filtering or normalization.

3.1.2 Region Growing

Region growing is a segmentation technique based on spatial intensity homogeneity. The process begins by selecting initial seed points within an image region, progressively aggregating neighboring pixels that share similar intensity values or texture characteristics. This technique iteratively expands the segmentation region based on predefined homogeneity criteria, typically intensity similarity or texture coherence [26]. Region growing is advantageous due to its intuitive implementation and ability to handle irregularly shaped regions effectively. However, it is sensitive to the selection of initial seed points and noise, which can result in leakage into adjacent regions if the criteria are inadequately defined.

3.1.3 Active Contour Models (Snake Algorithm)

Active contour models, widely known as "snakes," utilize energy-minimizing splines to delineate object boundaries dynamically. This technique involves an initial contour placed near the region of interest, evolving iteratively under the influence of internal forces (elasticity and rigidity constraints) and external forces (image gradient and intensity-based features) [38]. Snakes can effectively handle complex and irregular boundaries by adapting

flexibly to the shape of anatomical structures such as ventricles, tumors, and subcortical structures. A significant strength of active contour models is their ability to incorporate smoothness constraints, thereby providing highly precise boundary delineation essential for clinical tasks. Nevertheless, they require careful parameter tuning, initialization, and may fail if the initial contour is not appropriately placed near the desired boundaries, or if the image gradient is weak.

3.1.4 Random Walker Algorithm

The random walker algorithm is a probabilistic graph-based segmentation technique that interprets image pixels as nodes in a graph, with edges connecting neighboring pixels based on intensity similarity. Seed points representing labeled regions guide the algorithm, which calculates the probability that a random walker starting from an unlabeled pixel will reach a seed of a particular class (foreground or background) first. Pixels are assigned labels corresponding to the highest probability class, thereby yielding a segmentation [74]. The random walker algorithm is effective in handling noisy images and provides robust segmentation results with minimal sensitivity to initialization. However, computational cost is relatively high, especially for volumetric images, due to the iterative probability calculations, which can pose challenges in large-scale clinical datasets.

3.1.5 Mean Shift Clustering

Mean shift clustering is a non-parametric, iterative algorithm used to locate local maxima of density functions, effectively clustering data points in the feature space. For image segmentation, mean shift operates by treating pixels as points in a combined spatial-intensity feature space, iteratively shifting each point towards the mean of points within its local neighborhood until convergence to modes of the underlying distribution. This method does not assume prior knowledge of the number of clusters, making it flexible and data-driven. Mean shift clustering can effectively identify irregularly shaped clusters and is relatively robust to noise. However, its computational complexity is significant due to

iterative shifts and convergence checks, making it less efficient for high-dimensional or large-volume datasets common in medical imaging.

In summary, traditional computational segmentation methods are foundational for medical image analysis, particularly due to their interpretability and robustness in simpler clinical scenarios. However, the limitations in handling complex structures, sensitivity to noise and parameter tuning, and computational inefficiency in larger datasets highlight the need for advanced computational techniques such as machine learning and deep learning methods, which will be detailed in subsequent sections.

3.2 Machine Learning Approaches

Machine learning (ML) methods have been extensively utilized for medical imaging tasks due to their effectiveness in pattern recognition, classification, and segmentation. In the context of schizophrenia analysis, traditional ML algorithms primarily rely on carefully engineered features derived from neuroimaging data. Among these methods, Support Vector Machines (SVMs) [16] and Random Forests (RFs) [7] have emerged as prominent approaches due to their robust predictive performance and relative interpretability.

3.2.1 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are supervised learning algorithms widely used for binary classification tasks. SVMs identify an optimal hyperplane in a high-dimensional feature space, maximizing the margin between classes. This hyperplane serves as the decision boundary separating data points of different classes (an instance, schizophrenic versus healthy controls). SVMs have proven particularly effective in neuroimaging studies, especially when the number of available samples is relatively limited, due to their capability to manage high-dimensional input spaces and avoid overfitting through the use of kernel functions [3]. Kernel functions, such as linear, polynomial, and Radial Basis Function (RBF) [1], enable SVMs to map inputs into higher-dimensional spaces for enhanced class sepa-

ration. Despite their strong predictive capabilities, a significant limitation of SVMs is the need for manual feature extraction, which can introduce bias and reduce the generalizability of the model across diverse datasets.

3.2.2 Random Forests (RFs)

Random Forests (RFs) constitute an ensemble learning approach consisting of multiple decision trees trained independently on random subsets of features and samples. Each tree in the forest votes for a class, and the class with the majority votes is chosen as the prediction. RFs offer several advantages for medical image analysis, including robustness to noise, resilience against overfitting, and the ability to handle large feature sets and diverse data types effectively. This makes them suitable for tasks like the classification of structural and functional brain images to discriminate schizophrenia from healthy controls or other psychiatric disorders [74]. A critical strength of RFs is their inherent capability to estimate feature importance, thus enhancing interpretability in clinical contexts. However, similar to SVMs, RFs also rely on carefully selected and engineered features, limiting their scalability and the full exploitation of available data.

Both SVM and RF algorithms have demonstrated consistent performance in schizophrenia-related neuroimaging tasks, particularly when combined with sophisticated feature extraction techniques, such as voxel-based morphometry (VBM) and functional connectivity analysis. Nevertheless, the inherent limitation in scalability, manual feature dependency, and difficulty in capturing complex nonlinear patterns have prompted the exploration of advanced deep learning techniques discussed in subsequent sections. Such deep learning approaches, including convolutional neural networks (CNNs), transformers, and graph neural networks (GNNs), have shown potential in automating feature extraction, thereby achieving higher accuracy and improved generalization for schizophrenia detection.

3.3 Deep Learning Approaches

Deep learning methodologies have revolutionized medical imaging segmentation and classification tasks by enabling automated, feature extraction directly from imaging data, thus eliminating the necessity for explicit manual feature engineering. In neuroimaging studies, especially within schizophrenia research, Convolutional Neural Networks (CNNs), U-Net architectures, and Fully Convolutional Networks (FCNs) have emerged as essential computational tools due to their exceptional performance in precisely delineating intricate brain structures, offering considerable improvements in segmentation accuracy compared to traditional approaches[51].

3.3.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are hierarchical deep learning models designed to effectively extract spatial features from image data. CNN architectures typically include multiple convolutional layers, pooling layers, and fully connected layers, allowing the extraction of increasingly abstract and complex representations from the input images. Convolutional operations use kernels or filters that slide across input images to produce feature maps by computing local spatial correlations:

$$(f * g)(x,y) = \sum_i \sum_j f(i,j)g(x-i,y-j)$$

where f denotes the input image and g the convolutional kernel. CNNs efficiently capture spatial context and localized features, making them ideal for tasks like segmentation and classification of brain MRI images [45].

3.3.2 U-Net Architecture

The U-Net architecture is a specialized CNN variant designed explicitly for biomedical image segmentation. It features an encoder-decoder structure consisting of a contracting path (encoder) that captures context and spatial features and an expanding path (decoder)

that facilitates precise localization. Critically, U-Net employs skip connections as shown in Figure 3.1 between the encoder and decoder paths to preserve detailed spatial information, enhancing segmentation accuracy particularly at the boundaries [59]

The key mathematical component of U-Net is the skip-connection mechanism, defined as:

$$U_{decoder}(x) = f_{decoder}(U_{encoder}(x) \oplus S(x))$$

where $U_{encoder}(x)$ and $U_{decoder}(x)$ represent features from encoder and decoder, respectively, $S(x)$ denotes skip connections, and $\oplus$ indicates concatenation.

3.3.3 3D U-Net

3D U-Net [79] is an advanced extension of the original U-Net architecture, specifically designed to process volumetric medical imaging data such as MRI scans. Unlike traditional 2D U-Nets, which operate on slice-wise images and may lose spatial continuity across slices, 3D U-Net incorporates volumetric convolutions that preserve three-dimensional context. This enables the network to capture spatial dependencies and anatomical structures more effectively, which is particularly valuable in brain imaging tasks where continuity across slices is critical.

Formally, the operation of a 3D convolutional layer can be expressed as:

$$y = F_{conv3D}(x) = x * W + b$$

where $*$ denotes the 3D convolution, x represents the input volume, W the learned 3D kernel weights, and b the bias term. By extending convolutions to three dimensions, 3D U-Net allows the model to leverage volumetric features rather than relying solely on slice-based 2D context.

Empirical studies have shown that 3D U-Net significantly improves volumetric segmentation performance, yielding more accurate delineation of complex anatomical structures,

including small subcortical regions relevant to schizophrenia research [51]. This makes it a natural baseline for volumetric segmentation in medical imaging applications.

Limitations. Despite its strengths, 3D U-Net has several limitations that motivate the development of improved architectures:

- **Computational Complexity:** Processing volumetric data greatly increases memory and computational requirements. Training on full 3D MRI volumes often exceeds GPU memory limits, necessitating patch-based training strategies that may lose some global context.
- **Overfitting Risk:** Due to the high dimensionality of 3D data and relatively small medical imaging datasets, 3D U-Nets are prone to overfitting unless regularization or augmentation is carefully applied.
- **Boundary Precision:** While 3D U-Net captures volumetric features, it often struggles with fine-grained anatomical boundaries, producing segmentations that may "bleed" across structures. This is particularly limiting in psychiatric imaging, where precise delineation of subcortical regions such as the hippocampus, amygdala, basal ganglia, and thalamus are clinically important.
- **Generalization Challenges:** Variability in MRI acquisition protocols across sites (e.g., different scanners, resolutions) can reduce model robustness, highlighting the need for architectures that are invariant to acquisition differences.

These limitations provide the motivation for our proposed **BoRefAttnNet**, which integrates boundary refinement and attention mechanisms to improve segmentation precision, particularly around anatomical edges where 3D U-Net performance degrades.

3.3.4 Fully Convolutional Networks (FCNs)

Fully Convolutional Networks (FCNs) are specialized CNN architectures designed explicitly for dense pixel-wise segmentation tasks. Unlike traditional CNNs, FCNs eliminate

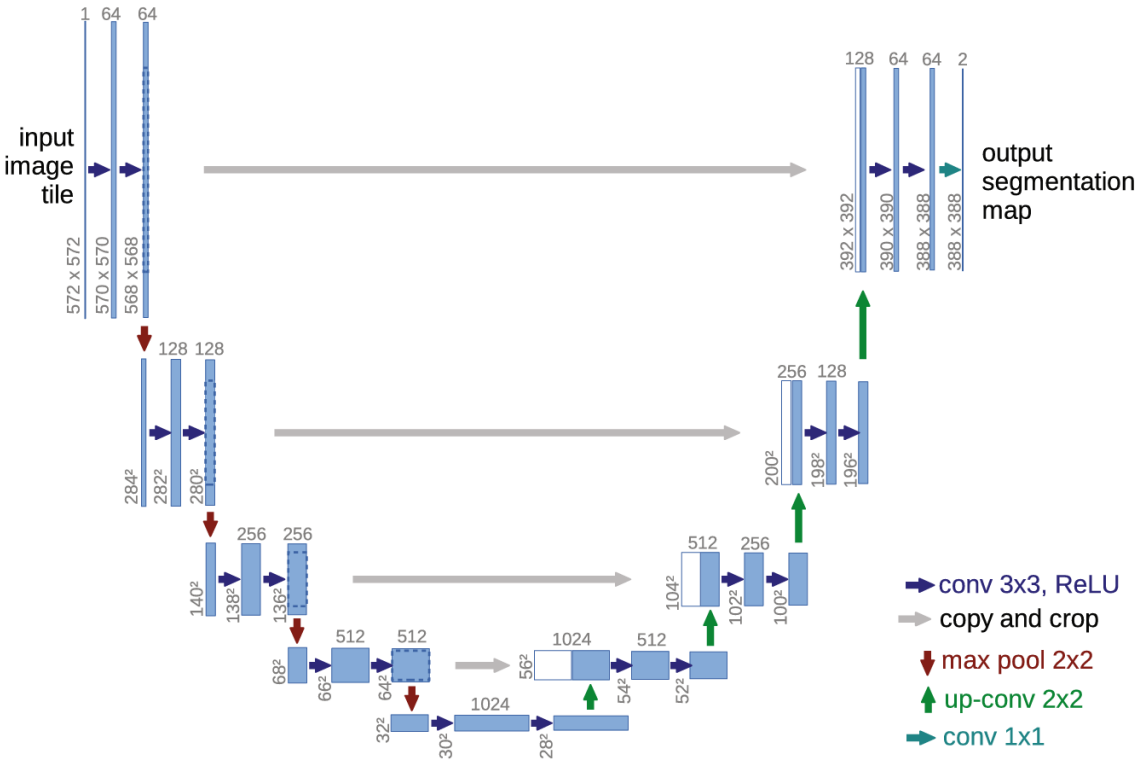


Figure 3.1: U-net architecture from Ronneberger et al. [59].

fully connected layers, allowing networks to handle images of arbitrary dimensions efficiently. FCNs provide dense outputs directly, preserving spatial relationships across the entire input image, and thus offer superior performance in medical image segmentation scenarios. Architectures such as SegNet and U-SegNet exemplify this category:

- **SegNet:** Utilizes an encoder-decoder architecture with pooling indices transferred from encoder to decoder to improve boundary accuracy and feature reconstruction.
- **U-SegNet:** Combines the structural advantages of SegNet and U-Net, incorporating skip connections and pooling indices to enhance spatial precision and boundary delineation [42].

3.4 Boundary-Aware Segmentation

Boundary-aware segmentation techniques, such as UDBRNet[28] and boundary-focused loss functions (Dice loss [77], boundary loss [39]), address critical limitations in standard

CNN-based methods concerning boundary precision. Studies comparing traditional U-Net and advanced boundary-aware networks such as UDBRNet consistently demonstrate improved segmentation performance when explicitly modeling boundary information and uncertainty. These comparative analyses highlight boundary-aware approaches as essential for precise anatomical delineation, especially for clinically sensitive analyses such as schizophrenia imaging.

Misalignment or incorrect boundaries can significantly affect clinical decisions, biomarker measurement accuracy, and therapeutic outcomes. In brain imaging, precise segmentation boundaries are particularly critical for delineating subtle anatomical alterations associated with schizophrenia, which can be key indicators of the disease[39]

Traditional segmentation methods and basic convolutional neural networks (CNNs) often lack explicit mechanisms for accurately capturing complex and irregular anatomical boundaries. Challenges such as noise, imaging artifacts, and intensity inhomogeneities further complicate boundary precision, necessitating advanced computational approaches tailored specifically for accurate boundary delineation.

3.5 Boundary-focused Loss Functions

Boundary-focused loss functions explicitly target the accuracy of segmentation boundaries by penalizing deviations from true boundary locations, thus improving segmentation precision significantly.

Boundary Loss Kervadec et al [39] proposed boundary loss directly penalizes boundary discrepancies between the predicted segmentation masks and ground truth contours. This loss function is mathematically expressed as:

$$\mathcal{L}_{Boundary} = \frac{1}{N} \sum_{p \in \partial G} d(p, \partial S)^2 + \frac{1}{M} \sum_{p \in \partial S} d(p, \partial G)^2 \tag{3.1}$$

where ∂G and ∂S represent ground truth and predicted segmentation boundaries, respectively; N and M denote the number of pixels in corresponding boundary sets; and $d(p, \partial S)$ represents the shortest Euclidean distance from point p to boundary set ∂S. Boundary loss effectively guides neural networks toward accurate alignment of predicted segmentation boundaries with the true anatomical contours[39]

Dice and Boundary-weighted Dice Loss The Dice loss, derived from the Dice similarity coefficient, measures overall segmentation overlap between prediction and ground truth masks, defined mathematically as:

$$\mathcal{L}_{Dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{3.2}$$

where X is the predicted segmentation mask and Y is the ground truth. Although Dice loss effectively captures global segmentation accuracy, boundary-weighted Dice loss improves it by incorporating boundary information to explicitly emphasize boundary pixels:

$$\mathcal{L}_{BW-Dice} = 1 - \frac{2\sum_i^N w_i X_i Y_i}{\sum_i^N w_i (X_i + Y_i)} \tag{3.3}$$

where w_i represents the boundary weighting factor applied to pixels near boundaries. This approach significantly enhances boundary delineation performance [66]

3.6 Vision Transformers

Transformer models, originally developed for natural language processing, have been effectively adapted for computer vision and medical imaging due to their powerful self-attention mechanism. Transformers capture complex spatial and temporal dependencies, providing significant advantages over traditional CNN architectures, particularly in neuroimaging analysis.

Self-Attention Mechanism Fundamentals The core of transformer architecture is the self-attention mechanism, enabling the model to capture long-range interactions by directly relating different elements within input data sequences. The self-attention operation is mathematically defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.4}$$

where Q, K, and V represent queries, keys, and values derived from input features, respectively, and d_k is the dimensionality of the key vectors. This mechanism allows transformer models to dynamically weigh and integrate information from all positions within the input data, overcoming the limitations of CNNs related to local receptive fields and fixed kernel sizes [14].

Advantages over CNNs in Medical Imaging Transformer-based approaches capitalize on self-attention mechanisms to capture long-range dependencies, significantly outperforming conventional CNN-based methods in schizophrenia classification by providing improved interpretability and robust detection performance [62]. Accordingly, Shehzad et al. [62] lists the following strengths of Transformers over CNNs

- **Long-range Dependencies:** Unlike CNNs, transformers directly capture global context and interactions across the entire image, essential for modeling long-distance dependencies characteristic of neuroimaging data.
- **Contextual Modeling:** Transformers naturally model context by attending simultaneously to multiple locations in images, facilitating the understanding of complex spatial-temporal relationships that CNNs may overlook due to localized convolutional operations.

3.7 Dynamic Spatial-Temporal Transformers

Dynamic spatial-temporal transformer models specifically extend transformer capabilities to handle sequential data over space and time, critical for dynamic brain MRI analysis.

Spatial-Temporal Transformers in Brain MRI Analysis Spatial-temporal transformers are designed to simultaneously leverage spatial information from imaging volumes and temporal dynamics across time-series data (e.g., fMRI sequences). These models capture rich spatio-temporal dependencies, allowing for enhanced diagnostic accuracy in identifying subtle neurological changes associated with schizophrenia. For instance, dynamic spatial-temporal transformers effectively integrate structural and functional neuroimaging data, yielding robust feature representations that significantly outperform traditional computational methods[62].

By explicitly modeling dynamic spatio-temporal dependencies, transformers can provide comprehensive insights into the neural mechanisms underlying schizophrenia, significantly surpassing conventional segmentation and classification approaches [62, 14]

3.8 Multimodal Fusion Techniques

Multimodal fusion strategies have become critical for improving diagnostic robustness and clinical accuracy by leveraging complementary information from multiple imaging modalities. Fusion of structural MRI (sMRI) and functional MRI (fMRI) significantly enhances schizophrenia detection capabilities compared to using single modalities, demonstrating clear clinical relevance in complex neuropsychiatric analyses.

Common fusion methods include feature-level concatenation, decision-level fusion, and hybrid fusion techniques using advanced computational approaches such as deep neural networks, transformers, and graph-based methods. Recent studies have notably emphasized transformer-based multimodal fusion approaches, which utilize self-attention mechanisms to dynamically weigh contributions from structural and functional modalities, thereby achiev-

ing superior diagnostic accuracy and interpretability [8]

While multimodal fusion strategies demonstrate improved diagnostic capabilities, they often encounter significant challenges. Strengths of current techniques include enhanced accuracy, robustness, and diagnostic interpretability. However, common limitations include computational complexity, the necessity for large datasets, sensitivity to noise, modality-specific artifacts, and difficulties in harmonizing data from diverse imaging sources. These challenges necessitate further methodological refinement and innovative computational architectures capable of effectively integrating multimodal imaging data.

3.9 Our Proposed Approaches

3.9.1 Limitations of Current Approaches

Traditional computational segmentation methods, including thresholding, region growing, and active contour models, remain foundational in medical imaging. Nevertheless, these techniques often face significant limitations, particularly concerning boundary accuracy, sensitivity to initialization parameters, and susceptibility to imaging noise [38]. Machine learning methods such as Support Vector Machines (SVMs) and Random Forests (RFs) provide effective classification capabilities but require extensive manual feature engineering, reducing their scalability and generalizability [3].

Recent deep learning approaches, notably Convolutional Neural Networks (CNNs), U-Net architectures (including 3D U-Net), Fully Convolutional Networks (FCNs), and hybrid models [29], have significantly improved segmentation accuracy by automatically extracting complex features directly from imaging data. Boundary-aware segmentation techniques, employing specialized loss functions, attention mechanisms, and refinement architectures (such as UDBRNet [28]), further enhance precision in segmentation tasks, particularly for intricate brain anatomy delineation.

Transformer models represent advanced computational techniques that effectively capture spatial-temporal relationships and global dependencies inherent in neuroimaging data.

Transformers, in particular, provide critical advantages over traditional CNN architectures, efficiently modeling long-range dependencies and contextual information crucial in brain imaging applications [62].

Despite these advancements, several computational challenges persist. High dimensionality, computational complexity, limited datasets, and significant inter-subject variability pose challenges in model training and validation. Achieving stable, interpretable, and clinically relevant results requires rigorous preprocessing, normalization, data harmonization, and robust model design, underscoring the importance of methodological rigor and innovation in computational neuroimaging research.

3.9.2 Motivations for Our Approach

Given these critical insights and existing computational challenges, the motivation for the current research emerges clearly. While significant progress has been made in leveraging advanced segmentation models, significant gaps remain in achieving boundary precision, interpretability, and robustness, especially in multimodal and spatial-temporal contexts. The necessity for enhanced boundary-aware segmentation methodologies and sophisticated modeling of dynamic spatial-temporal dependencies forms the core motivation of this research.

Specifically, this thesis addresses the following critical needs:

1. **Enhanced Boundary Precision:** Developing advanced segmentation methods explicitly optimized for accurate boundary delineation to precisely identify subtle anatomical alterations associated with schizophrenia.

2. **Spatial-Temporal Integration:** Employing transformer-based architectures capable of effectively modeling complex spatial and temporal brain connectivity dynamics, essential for accurately capturing schizophrenia pathology and progression.

3. **Multimodal Fusion:** Implementing computational frameworks that integrate struc-

tural and functional MRI modalities to leverage complementary strengths and improve diagnostic accuracy, robustness, and clinical interpretability.

Ultimately, this research aims to push the boundaries of existing computational methodologies, addressing identified limitations, enhancing the interpretability and diagnostic capability of schizophrenia analysis, and contributing significantly to advancing clinical decision-making and patient outcomes.

Chapter 4

BoRefAttnNet: Boundary-Refined Attention Network

Accurate boundary delineation of brain anatomical structures in 3D medical image segmentation remains a critical challenge, particularly for anatomically subtle regions with boundaries vital for clinical and neuroscientific insights. In this chapter, we present our variant Boundary-Refined Attention Network *(BoRefAttnNet)*, a boundary-refined 3D U-Net variant that integrates multiscale boundary attention modules into the decoder path. In place of standard skip-connections, the modules in our model produce explicit boundary activation maps that selectively emphasize anatomically relevant edges while suppressing background noise. We evaluate BoRefAttnNet on the Analysis of Functional NeuroImages (AFNI)-processed brain magnetic resonance imaging of patients (acquired through the SchizConnect license from the Center for Biomedical Research Excellence (COBRE) dataset) that are pre-processed for skull stripping and FastSurfer parcellations. Focusing on five key subcortical structures (hippocampus, lateral ventricles, amygdala, basal ganglia, and thalamus) as our class labels, our model trains under two loss configurations, cross-entropy alone and cross-entropy with boundary penalty. A set of experiments show our BoRefAttnNet model substantially outperforms the current conventional 3D U-Net baselines, resulting in more precise segmentations of small or complex structures.

4.1 Boundary-Refined Attention Network (BoRefAttnNet)

Despite significant advances, existing boundary-aware segmentation models have critical limitations, which include issues, such as computational complexity and insufficient attention to subtle anatomical boundaries, especially in brain MRI. To this end, our proposed approach incorporates explicit boundary attention blocks at multiple decoder scales and uniquely refines boundary delineation through targeted attention mechanisms. By selectively enhancing boundary-relevant features without significantly increasing computational overhead, our approach presents a clear technical advancement suited to the challenging task of precise anatomical segmentation in volumetric MRI.

Although modern 3D U-nets have achieved the strong volumetric encoding and multiscale analysis, conventional encoder-decoder pipelines often struggle with accurately localizing object boundaries—particularly in highly irregular or subtle brain regions, such as the hippocampus or amygdala. In these cases, the decoder skip-connections transfer high-resolution features but lack explicit cues to emphasize tissue interfaces, which leads to blurred or imprecise segmentations. We propose *Boundary-Refined Attention Network BoRefAttnNet*, a novel 3D U-Net variant that integrates *multi-scale Boundary Attention Modules* (BAM) directly into its decoder stages. These modules produce *boundary activation maps* that highlight transition zones between anatomical structures, gating the decoder features to preserve edges while suppressing homogeneous or uninformative regions. By design, BoRefAttnNet not only preserves the benefits of the encoder-decoder framework, but also focuses computational resources on boundary delineation, an aspect especially vital for accurate neuroimaging assessments in schizophrenia.

The primary objectives of BoRefAttnNet include the following.

1. *Boundary Accuracy*: Explicitly model boundaries to improve segmentation precision at structural interfaces.

2. *Multi-scale Contextual Learning*: Integrate boundary attention at multiple decoder resolutions for comprehensive boundary modeling.

3. *Computational Efficiency*: Maintain comparable computational demands to baseline 3D U-Net architectures, allowing scalable deployment.

4.1.1 Model Architecture

As illustrated in Figure 4.1, the encoder follows a standard 3D U-Net downsampling pathway. Each decoder stage integrates a Boundary Attention Module (BAM) that refines skip-connected features by emphasizing boundary regions. At every stage, the model produces boundary logits, with the final boundary map taken from the last decoder stage. A 1×1 convolution on the final decoder output yields the per-voxel segmentation logits.

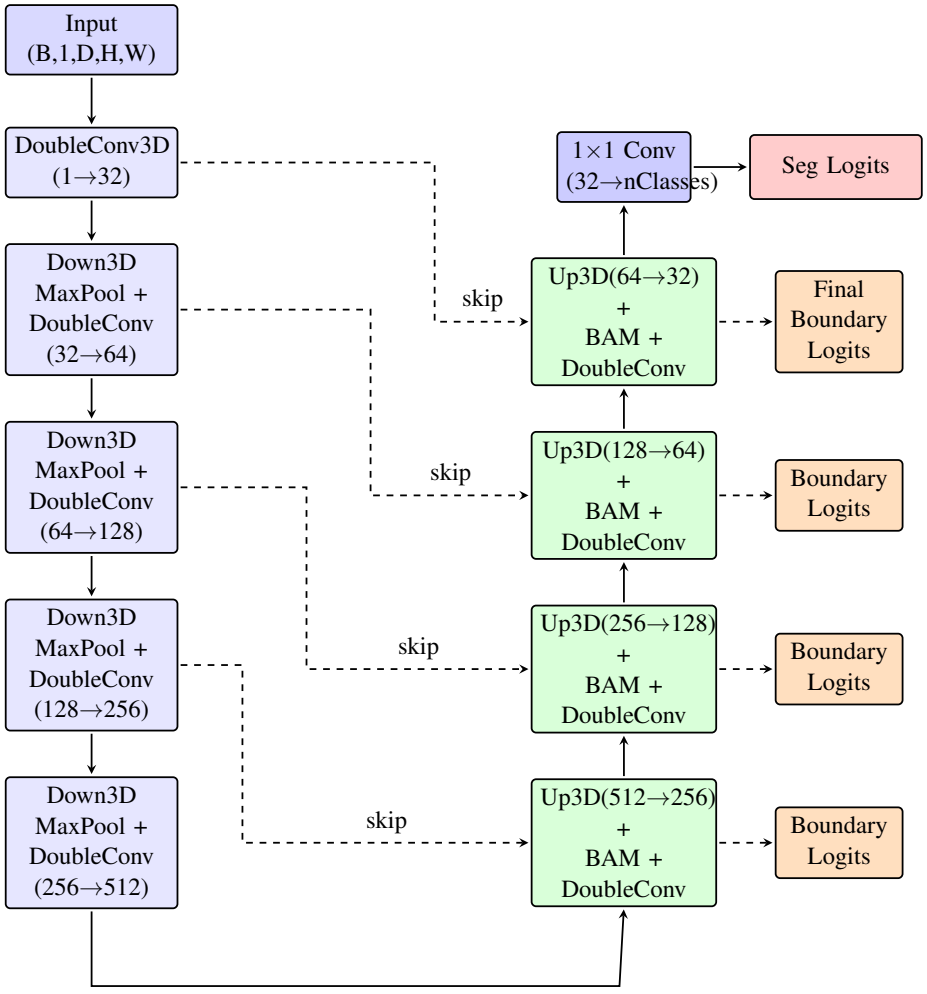


Figure 4.1: Schematic of BoRefAttnNet showing the encoder-decoder architecture with boundary attention modules (BAM) integrated at each decoder stage.

BoRefAttnNet inherits its overall structure from the original 3D U-Net [59] but aug-

ments the decoder itself with Boundary Attention Modules (BAM) at multiple resolutions.

Formally, let

$$\mathcal{X} \subset \mathbb{R}^{C\times D\times H\times W} \tag{4.1}$$

be a given input *T1-weighted MRI* volume—a standard anatomical sequence that provides high soft-tissue contrast, helping to distinguish gray matter, white matter, and subcortical structures—where $C = 1$ for channel, showing single modality (MRI collected as grayscale images), D is the depth which in volumetric MRI, this corresponds to the number of slices along the z-axis., H is the height which represents the number of pixels along the y-axis of each slice, and W the width is the number of pixels along the x-axis of each slice. The network consists of the following modules.

1. A 3D *Encoder* $\mathcal{E}$ that down-samples features hierarchically.

2. A series of Decoder Stages $\{\mathcal{D}_l\}$, where each stage l merges encoder features with upsampled decoder features.

3. Boundary Attention Modules inserted at each decoder stage to create boundary activation maps $\mathcal{B}^{(l)}$. These maps selectively weigh the skip-connected features, emphasizing regions that exhibit boundary transitions.

The function-level segmentation output description is as follows.

$$\mathcal{Y} \;=\; \mathcal{D}\big(\mathcal{A}^{(1)}(f_e^{(1)}, f_d^{(1)}),\; \ldots,\; \mathcal{A}^{(L)}(f_e^{(L)}, f_d^{(L)})\big) \tag{4.2}$$

where $\mathcal{A}^{(l)}$ denotes the boundary attention module (or boundary activation map) at stage l, $f_e^{(l)}$ are encoder features, and $f_d^{(l)}$ are upsampled decoder features, where $L = 4$ in Figure 4.1.

Each boundary attention module aims to isolate local transitions between tissues, for instance, between hippocampus and ventricle, by learning an auxiliary boundary activation

map $\mathcal{B}^{(l)}$. This boundary map gates the concatenated encoder and decoder features $\left(f_e^{(l)} \oplus f_d^{(l)}\right)$, compelling the network to assign higher weights to boundary voxels.

$$B^{(l)} = \sigma\left(\phi_{1\times1\times1}\left(\delta\left(\phi_{3\times3\times3}(f_e^{(l)} \oplus f_d^{(l)})\right)\right)\right) \tag{4.3}$$

where $\phi_{3\times3\times3}$ and $\phi_{1\times1\times1}$ denote convolutional operations, δ represents Group Normalization [72] followed by ReLU [49] activation, σ is the sigmoid function generating the attention map, and $\oplus$ denotes concatenation.

The attention-gated decoder features are subsequently computed by element-wise multiplication:

$$f_{attn}^{(l)} = B^{(l)} \odot (f_e^{(l)} \oplus f_d^{(l)}) \tag{4.4}$$

where $\odot$ denotes Hadamard (element-wise) product [33]. This gating mechanism selectively emphasizes voxels along anatomical edges, yielding sharper segmentation contours.

4.1.2 Encoder: Multi-Scale Feature Extraction

The encoder leverages hierarchical 3D convolutional blocks to extract multi-scale spatial features critical for accurate boundary delineations. Each encoder block employs a two-stage convolutional sequence (DoubleConv3D), including $3 \times 3 \times 3$ convolutions, Group Normalization, and ReLU activations, followed by downsampling via MaxPooling.

Let the input volume be $x \in \mathcal{X}(\in \mathcal{R}^{C\times D\times H\times W})$, where $C = 1$ for single-channel MRI data. Encoder operations at each stage $l \in \{1,2,3,4\}$ are formally described as follows.

$$f_e^{(0)} = x, \quad f_e^{(l)} = \text{Down3D}\big(\text{MaxPool3D}(f_e^{(l-1)})\big) \tag{4.5}$$

Each *Down3D* consists of two consecutive $3 \times 3 \times 3$ convolutions, each followed by Group Normalization and ReLU activation, then a max-pooling step (see Table 4.1). After four such stages ($l = 1,\ldots,4$), we obtain progressively deeper features $f_e^{(4)}$ capturing broad context.

Table 4.1: BoRefAttnNet Encoder Specifications.

Stage	Input Ch.	Output Ch.	Kernel	Norm	Downsample
1	1	32	$3 \times 3 \times 3$	GroupNorm (8)	MaxPool3D
2	32	64	$3 \times 3 \times 3$	GroupNorm (8)	MaxPool3D
3	64	128	$3 \times 3 \times 3$	GroupNorm (8)	MaxPool3D
4	128	256	$3 \times 3 \times 3$	GroupNorm (8)	MaxPool3D

4.1.3 Decoder: Boundary-Aware Upsampling and Multi-Scale Attention

A short *bottleneck* block converts the deepest encoder output $f_e^{(4)}$ into the initial decoder feature, which we denote $f_d^{(\text{bot})}$. Then, each decoder stage $l \in \{4,3,2,1\}$:

- Upsamples a deeper decoder feature $f_d^{(l+1)}$ or $f_d^{(\text{bot})}$ (for $l = 4$) to match spatial resolution;
- Concatenates it with the encoder feature $f_e^{(l)}$ (a skip connection);
- Passes the concatenated volume through a *Boundary Attention Module* (BAM) block, described below;

Formally, let us define:

$$f_d^{(\text{bot})} = \text{Bottleneck}\big(f_e^{(4)}\big)$$

as the output from the deepest layer. Then, for each decoder stage $l \in \{4,3,2,1\}$ we have:

$$f_d^{(l)} = \text{UpConv3D}\big(f_d^{(l+1)}\big) \tag{4.6}$$

$$\big(\text{or } f_d^{(\text{bot})} \text{ if } l = 4\big) \tag{4.7}$$

$$x^{(l)} = \text{Concat}\big(f_d^{(l)}, f_e^{(l)}\big) \tag{4.8}$$

$$x_{\text{refined}}^{(l)}, B^{(l)} = \text{BAM}\big(x^{(l)}\big) \tag{4.9}$$

where $B^{(l)}$ is the boundary logits at stage l. This gating mechanism in BAM selectively emphasizes edge-critical voxels and yields sharper segmentation contours. Finally, a $1 \times$

1×1 convolution on the last decoder output produces the per-voxel segmentation logits.

4.1.4 Boundary Attention Module (BAM)

Each BAM is anchored around a small set of learnable convolutional layers that predict a 3D boundary activation map. Let $x \in \mathcal{X}(\subset \mathcal{R}^{C\times D\times H\times W})$ be the skip-concatenated feature map; the boundary logits B_{logits} are produced by our BAM module as follows. The whole architecture of our BAM module is shown in Figure 4.2.

$$x_{\text{int}} \;=\; \text{ReLU}\big(\text{GN}\big(\text{Conv}_{3\times3\times3}(x)\big)\big) \tag{4.10}$$

$$B_{\text{logits}} \;=\; \text{Conv}_{1\times1\times1}(x_{\text{int}}) \tag{4.11}$$

Applying a sigmoid activation $\sigma(\cdot)$ to the logits yields a continuous boundary attention map A_{boundary}, which in turn gates the feature map x by element-wise multiplication:

$$x_{\text{refined}} \;=\; x \odot A_{\text{boundary}} \tag{4.12}$$

This multi-scale attention mechanism enhances boundary clarity by emphasizing edge-critical voxels across multiple resolution scales, significantly improving segmentation accuracy for intricate anatomical boundaries.

Algorithm 1 provides the pseudocode of our BAM module, a $3 \times 3 \times 3$ conv with GN and ReLU produces intermediate features, followed by a $1 \times 1 \times 1$ conv for boundary logits. The B_{logits} passes through a sigmoid σ, generating an attention map that gates the original input x by element-wise multiplication, yielding x_{refined}. This gating localizes transitions along edges, mitigating the risk of over-smoothing that standard skip-connections often exhibit.

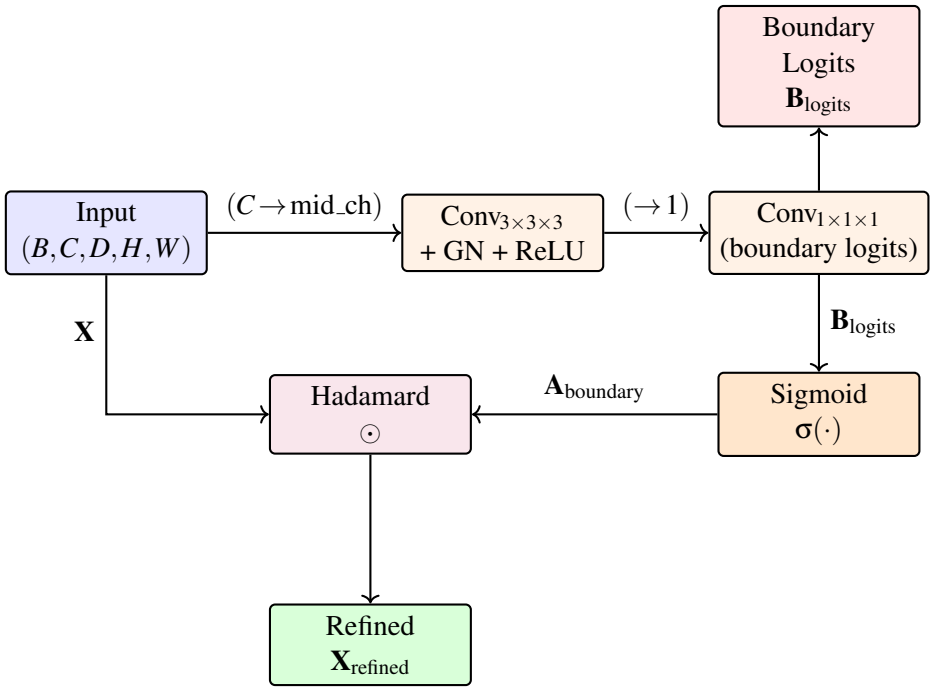


Figure 4.2: Architecture of the Boundary Attention Module (BAM).

Algorithm 1 Boundary Attention Module (BAM)

Input: Feature map $x \in (\mathcal{X} \in \mathbb{R}^{B\times C\times D\times H\times W})$

Output: Refined feature map x_{refined}, boundary logits B_{logits}

$x_{\text{int}} \leftarrow \text{ReLU}\big(\text{GN}\big(\text{Conv}_{3\times3\times3}(x)\big)\big)$

$B_{\text{logits}} \leftarrow \text{Conv}_{1\times1\times1}\big(x_{\text{int}}\big)$

$A_{\text{boundary}} \leftarrow \sigma(B_{\text{logits}})$

$x_{\text{refined}} \leftarrow x \odot A_{\text{boundary}}$

return $x_{\text{refined}}, B_{\text{logits}}$

4.1.5 Boundary-Aware Loss Formulation

In addition to conventional volumetric Dice and Cross-Entropy[66] metrics, BoRefAttnNet leverages a hybrid *boundary-focused loss*, integrating volumetric accuracy (Dice Loss) with explicitly formulated boundary precision to further constrain edge quality.

Dice Loss quantifies the overlap between predicted mask P and ground truth G, defined

as:

$$L_{\text{Dice}}(P,G) = 1 - \frac{2\sum_i P_i G_i}{\sum_i P_i^2 + \sum_i G_i^2} \tag{4.13}$$

The boundary loss term [39] explicitly penalizes misalignments at object contours by comparing predicted boundaries to ground truth edges using gradient approximations as:

$$L_{\text{Boundary}} = \frac{1}{N} \sum_{x \in \Omega} w_x \left|\nabla P(x)\right|^2 \left(P(x) - G(x)\right)^2 \tag{4.14}$$

where w_x is the voxel-wise boundary weighting factor, $\nabla P(x)$ denotes spatial gradients approximated across each dimension and Ω denotes the set of all (x, y, z) voxel coordinates in the 3D volume, and $N = |\Omega|$ is the total number of voxels. In other words, we sum over every voxel location in the image lattice when computing the boundary loss.

We evaluate our baseline 3D U-Net and the proposed *BoRefAttnNet* on a 6-class subcortical segmentation task derived from FastSurfer [31, 32, 22, 21] labeled brain MRI data.

4.2 Data Preparation and Training

4.2.1 Data Preparation

Our experiments utilize structural MRI data from the publicly available *Center for Biomedical Research Excellence* (COBRE) [2] dataset obtained via the *SchizConnect platform* [1] [2]. Specifically, our analyses focused on T1-weighted structural MRI scans, acquired using a Siemens 3T TIM Trio scanner using a 12-channel head coil. Images were acquired using a multi-echo MPRAGE sequence with the following parameters: TR = 2530 ms, TE = 1.64, 3.5, 5.36, 7.22, 9.08 ms (multi-echo), TI = 1.2 s, flip angle = 7°, voxel size = 1 mm³ isotropic, field of view (FOV) = 256 mm, and acquisition matrix of 256×256×192 voxels. Figure 4.3 provides a sample MRI data, visualizing the human brain anatomy from different perspectives—sagittal, coronal, and axial—providing a comprehensive overview of structural abnormalities such as enlarged ventricles or cortical thinning, commonly observed in

[1] COBRE: `https://pubmed.ncbi.nlm.nih.gov/28812221/`.
[2] SchizConnect platform: `http://schizconnect.org/`.

patients with schizophrenia.

Explanation of Parameters. TR (repetition time) specifies the time between successive excitations of the same slice, influencing image contrast. TE (echo time) is the time between excitation and signal readout, controlling sensitivity to tissue relaxation. TI (inversion time) defines the delay between an inversion pulse and signal acquisition, crucial for contrast in MPRAGE sequences. The flip angle determines the degree of proton rotation by the RF pulse, and the FOV defines the spatial extent of the imaging volume. These acquisition parameters together provide high-resolution T1w images suitable for precise volumetric and morphological analysis.

Subjects and Class Labels. The dataset included 174 subjects: 88 diagnosed with schizophrenia and 86 healthy controls. For the purposes of segmentation and classification, these two groups form the binary class labels used in this study. In downstream segmentation experiments, we further refined class labels to specific anatomical regions of interest (ROIs), including the hippocampus, amygdala, lateral ventricles, basal ganglia, and thalamus, consistent with prior studies reporting schizophrenia-related abnormalities.

Data Split. The dataset was divided into training (70%), validation (15%), and testing (15%) subsets to ensure rigorous model evaluation while preventing data leakage. This corresponds to approximately 122 subjects for training, 26 for validation, and 26 for testing, stratified across schizophrenia and control groups. This balanced split ensures that both diagnostic categories are represented across all subsets, providing a reliable benchmark for segmentation and classification tasks in schizophrenia-related neuroimaging studies [69, 52]. Previous works indicate significant volumetric alterations in regions such as the hippocampus, amygdala, and lateral ventricles among patients with schizophrenia [50, 71]. Therefore, the experiment of our segmentation performance evaluation specifically targets these anatomically and clinically critical regions.

Initial preprocessing of the data involves skull stripping using the *Analysis of Functional*

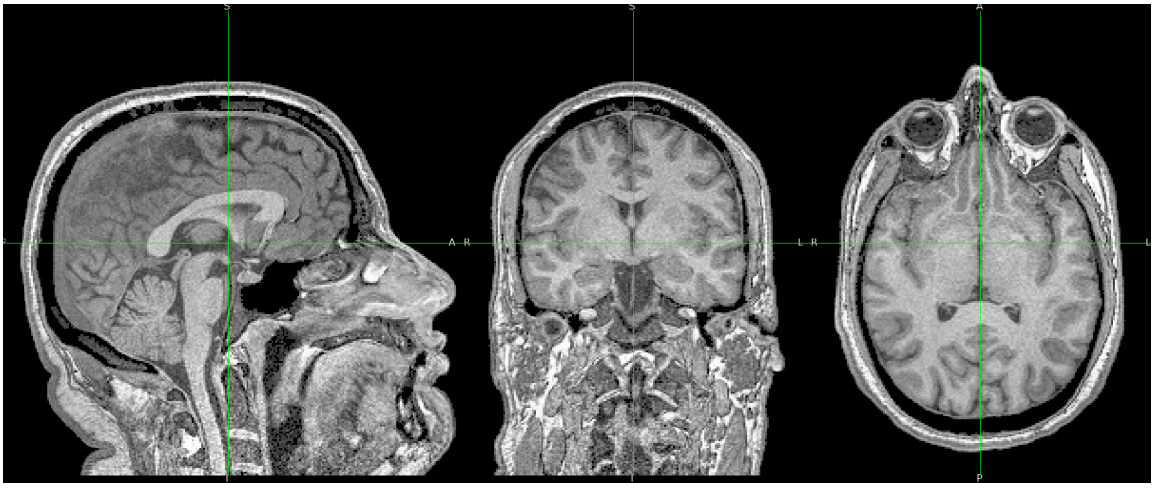


Figure 4.3: Orthogonal view of a T1-weighted MRI showing the Sagittal, Coronal, and Axial slices

NeuroImages AFNI[3] [18, 17] *sswarper2*. This step isolates brain tissue from surrounding skull and extracerebral tissues, significantly reducing noise for downstream tasks.

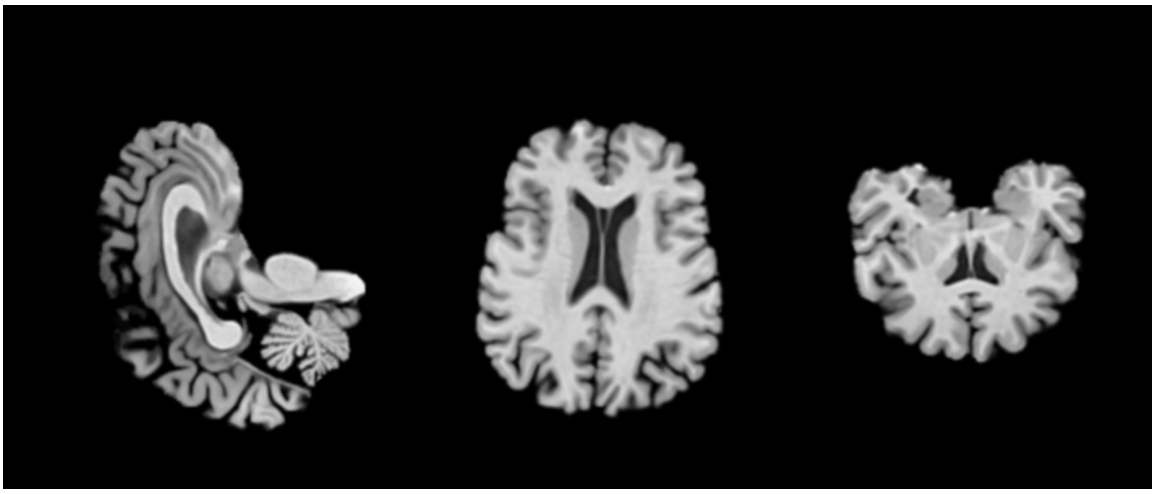


Figure 4.4: Sagittal, Coronal, and Axial slices view of an input MRI for model training.

After skull-stripping, intensity normalization is applied, standardizing each scan to zero mean and unit variance. This means that the average voxel intensity across the entire volume is shifted to zero (centering), and the variance of voxel intensities is scaled to one (standardization). The skull-stripped volumes undergo further normalization using *Fast-*

[3]AFNI: `https://afni.nimh.nih.gov`.

Surfer[4] to generate outputs for each subject. Our training inputs as shown in Figure 4.4 are T1-weighted MRIs which have been preprocessed, skull stripped, and are passed through FastSurfer for further normalization. We adopt FastSurfer-generated label volume shown in Figure 4.5 that combine multiple cortical and subcortical parcellations.

Figure 4.5: FastSurfer generated multi-class label of subject MRI in Sagittal, Coronal, and Axial slice view.

We refer to FastSurfer's Color LUT to isolate and merge the data into six (6) *clinically relevant subcortical* classes, namely *Background*, *Hippocampus*, *Lateral Ventricles*, *Amygdala*, *Basal Ganglia*, *Other structures*.

4.2.2 Computational Environment

Experiments are conducted using high-performance computing resources provisioned on Google Cloud Platform (GCP) [5] , specifically configured to optimize training and evaluation of deep neural networks:

- *Compute Instance:* Google Cloud VM instance type *g2-standard-48*.
- *Hardware Specifications:* Equipped with 48 virtual CPUs (vCPUs), 192 GB of RAM, and four NVIDIA L4 GPUs (32 GB VRAM each), ensuring high computational

[4]FastSurfer: `https://github.com/Deep-MI/FastSurfer`.
[5]Console: `https://cloud.google.com`.

throughput and efficient parallelization.

- *Storage:* Solid-state drives (SSDs) provided rapid data access and significantly minimized I/O latency.

4.2.3 Software and Libraries

The software stack and tools leveraged for conducting the experiments included:

- *Deep Learning Framework:* PyTorch 2.6 [6] for model development, training, and evaluation.
- *Data Handling Libraries:* Nibabel [7] for neuroimaging data manipulation and preprocessing.
- *Visualization Tools:* FSL [8] for interpreting model attention and outputs.
- *Version Control:* Git and GitHub [9] ensured reproducibility and maintained version consistency across experiments.

The following models and loss configurations are implemented.

- Baseline 3D U-Net (Cross Entropy loss-only).
- Baseline 3D U-Net with Boundary Penalty (Cross Entropy loss+boundary loss).
- BoRefAttnNet with multi-scale boundary attention (Cross Entropy loss-only).
- BoRefAttnNet with multi-scale boundary attention and boundary-weighted loss (Cross Entropy loss+boundary loss).

A patch-based approach is adopted from *kamnitasas et al.* [37] with patch size 128^3, stride 64 and with independent data augmentation per patch. Adam optimizer [40], initial learning rate $1e^{-4}$, and early stopping [56] criteria based on validation loss are employed.

[6] Pytorch: `https://pytorch.org`.
[7] Nibabel: `https://nipy.org/nibabel/`.
[8] FSLeyes: `https://fsl.fmrib.ox.ac.uk/`.
[9] Github: `https://github.com/`.

4.2.4 Evaluation Metrics

Three (3) standard metrics are used to evaluate segmentation accuracy, including *Dice Similarity Coefficient* (Dice), *Hausdorff Distance* (HD), and *Average Surface Distance* (ASD).

To illustrate the efficacy and interpretability of our proposed BoRefAttnNet, in our experiments we also visually examine the intermediate boundary attention maps generated at various decoder stages. These boundary attention maps explicitly highlight the anatomical boundaries critical for accurate segmentation, particularly in challenging structures involved in schizophrenia analysis.

The boundary attention maps are generated using the boundary logits $B^{(l)}$ at each decoder stage l, which are subsequently passed through a sigmoid activation function to obtain a normalized boundary attention map [61].

$$A^{(l)} = \sigma(B^{(l)}) = \frac{1}{1 + e^{-B^{(l)}}} \tag{4.15}$$

where $A^{(l)}$ represents the boundary attention map at decoder stage l.

4.3 Experiments

Experiment results are summarized in Table 4.2. As shown in the table, the segmentation outcomes comparing baseline 3D U-Net and our BoRefAttnNet variants demonstrate improved delineation accuracy using BoRefAttnNet, particularly around subcortical boundaries. On inference test, the CE-only model achieved a macro-average Dice of 0.9655, Hausdorff distance (HD) of 4.57 mm, and average surface distance (ASD) of 0.18 mm. Augmenting the loss with our introduced boundary penalty shows further improvement of the performance of our proposed model, raising Dice to 0.9691 and reducing HD and ASD to 3.08 mm and 0.16 mm, respectively.

In addition to reporting the *Macro Dice*, i.e., the unweighted average of the per-class Dice values across all six classes, Table 4.3 details each class's Dice. Notably, smaller

Table 4.2: Quantitative comparison of segmentation results on the test set.

Method	**Loss**	**Dice (macro)**	**HD (mm)**	**ASD (mm)**
3D U-Net	CE-only	0.9425	7.90	0.29
3D U-Net	CE+boundary	0.9540	6.45	0.23
BoRefAttnNet	CE-only	0.9655	4.57	0.18
BoRefAttnNet	CE+boundary	**0.9691**	**3.08**	**0.16**

or more irregular classes such as the amygdala see larger relative gains from boundary attention.

Table 4.3: Per-class Dice for BoRefAttnNet (CE+boundary). BG=background, Hipp=hippocampus, Vent=ventricle, Amyg=amygdala, BasGang=basal ganglia, Thal=thalamus.

Class	BG	Hipp	Vent	Amyg	BasGang	Thal
Dice	0.9999	0.9616	0.9698	0.9462	0.9702	0.9688

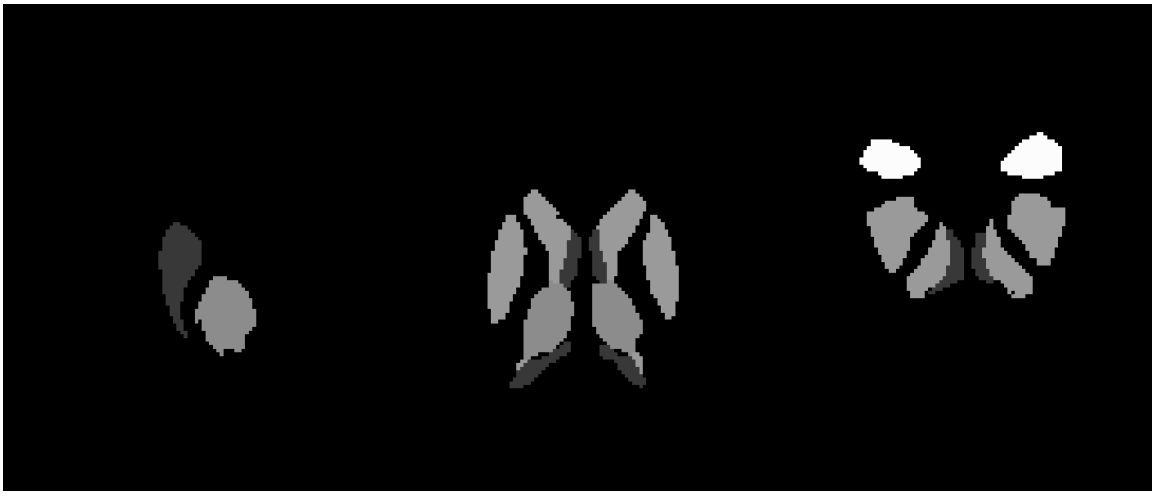


Figure 4.6: Qualitative segmentation results on a test subject.

4.4 Summary

In our experiments, the proposed BoRefAttnNet demonstrates significant improvements in boundary delineation compared to the baseline 3D U-Net architecture. Figure 4.6 presents

representative slice overlays of the predicted segmentations produced by BoRefAttnNet on a test subject. These visualizations highlight the model's enhanced capability in delineating key subcortical structures such as the amygdala (the two small lateral regions located inferiorly on the left and right sides), the hippocampus (elongated medial structures adjacent to the amygdala), the basal ganglia and thalamus (larger central clusters in the middle of the slice), and the lateral ventricles (visible as superior bilateral regions, appearing brighter due to their distinct boundaries), particularly capturing intricate boundaries along anatomically complex interfaces like the hippocampus–amygdala junction.

Quantitatively, BoRefAttnNet significantly improves segmentation metrics, notably the Hausdorff Distance (HD) and Average Surface Distance (ASD), which specifically quantify boundary accuracy. The model achieves a reduction in HD to 3.08 mm and ASD to 0.16 mm, outperforming the traditional 3D U-Net. These metrics underscore BoRefAttnNet's capacity to precisely delineate anatomical boundaries, essential for accurate neuroanatomical assessments critical in clinical decision-making and schizophrenia diagnosis.

Furthermore, leveraging multi-scale Boundary Attention Modules (BAM), the model successfully captures hierarchical spatial information, resulting in superior performance across multiple anatomical structures. Particularly notable improvements are observed for smaller, more irregularly shaped regions, such as the amygdala, which traditionally pose significant challenges for standard segmentation architectures.

The improved boundary precision not only demonstrates methodological advancement but also holds significant potential for clinical applications by enabling more accurate morphometric analyses and facilitating early detection of subtle anatomical alterations associated with schizophrenia pathology. Thus, BoRefAttnNet represents a meaningful advancement in boundary-aware medical image segmentation, demonstrating robust potential for clinical diagnostics and neuroimaging research.

Chapter 5

Schizophrenia Detection using Dynamic Spatial-Temporal Transformer Model (DySTTM)

This chapter focuses on the critical task of detecting schizophrenia using neuroimaging data. Building on Chapter 3, which actualizes precise boundary-aware segmentation of structural MRI, in this chapter, we will incorporate functional MRI (fMRI) data together with structural MRI (sMRI) data, capitalizing on their complementary temporal and spatial information to better detect schizophrenia. The proposed methodology utilizes a Spatial-Temporal Transformer architecture specifically designed to leverage dynamic interactions between structural and functional modalities, addressing existing methodological limitations.

5.1 Motivation for Spatial-Temporal Transformers

Existing schizophrenia detection models primarily rely on either structural MRI (sMRI) or functional MRI (fMRI) independently, limiting their capability to fully leverage the intricate interplay between anatomical disruptions and functional connectivity alterations. Structural MRI is proficient in identifying anatomical abnormalities, particularly volumetric differences in critical brain regions such as the hippocampus, amygdala, and basal ganglia, which have been robustly associated with schizophrenia pathology [2, 8]. In contrast, functional MRI provides insights into disrupted neural connectivity patterns and abnormal temporal dynamics across essential brain networks, including the Default Mode Network

(DMN) and the Salience Network [46, 47].

However, traditional analytical frameworks, often based on simplified multivariate approaches or unimodal analyses, inadequately represent the dynamic spatial-temporal interactions between these structural and functional domains [10]. As a result, critical spatial-temporal information can be overlooked, negatively impacting detection accuracy and reducing interpretability regarding underlying disease mechanisms [48].

To address these limitations, we propose adopting a Spatial-Temporal Transformer model explicitly designed to leverage:

1. *Spatial Attention Mechanisms:* Capturing anatomical localization and structural specificity through detailed segmentation masks derived from the structural MRI.

2. *Temporal Attention Mechanisms:* Modeling dynamic fluctuations and temporal coherence of neuronal activity from resting-state fMRI [12, 47].

The Transformer architecture would inherently support capturing long-range dependencies and interactions, making it especially suitable for modeling complex brain connectivity and structural-functional associations [14]. Furthermore, attention mechanisms within the Transformers allow explicit interpretation of feature importance, enhancing clinical interpretability [13, 62]. Thus, integrating structural and functional data through a Spatial-Temporal Transformer would not only provide superior classification performance but also enable deeper insights into the neurobiological underpinnings of schizophrenia [10]. It is expected that our work significantly advances the diagnostic potential of neuroimaging-based biomarkers [36].

5.2 Dynamic Spatial-Temporal Transformer Model

The Dynamic Spatial-Temporal Transformer Model (DySTTM) leverages a dual-stream architecture to effectively integrate and exploit structural and functional MRI modalities. This architecture distinctly manages spatial (anatomical) and temporal (functional) infor-

mation, integrating them through advanced multi-head attention mechanisms. Figure 5.1 provides a schematic diagram of the proposed dual-stream Transformer.

5.2.1 Two-Stream Architecture

Our approach, illustrated in Figure 5.1 comprises two separate streams dedicated to structural MRI and functional MRI data. This approach ensures specialized processing and maximizes feature extraction for each modality.

Structural Stream

The structural stream incorporates anatomical features derived from segmentation outputs obtained via BoRefAttnNet (Chapter 4). Specifically, the structural stream captures volumetric and morphological metrics, including regional volume, cortical curvature, and thickness. These features, denoted as $\mathbf{X}_s \in \mathbb{R}^{N \times D}$, where N represents the number of segmented anatomical regions and D the dimensionality of extracted structural features, are processed through spatial attention blocks, mathematically formulated as:

$$\text{Attention}(Q_s, K_s, V_s) = \text{softmax}\left(\frac{Q_s K_s^\top}{\sqrt{d_k}}\right) V_s$$

where Q_s, K_s, V_s represent query, key, and value matrices, respectively, constructed from structural embeddings [68]. Spatial attention enables the model to recognize crucial anatomical interdependencies relevant to schizophrenia pathology.

Functional Stream

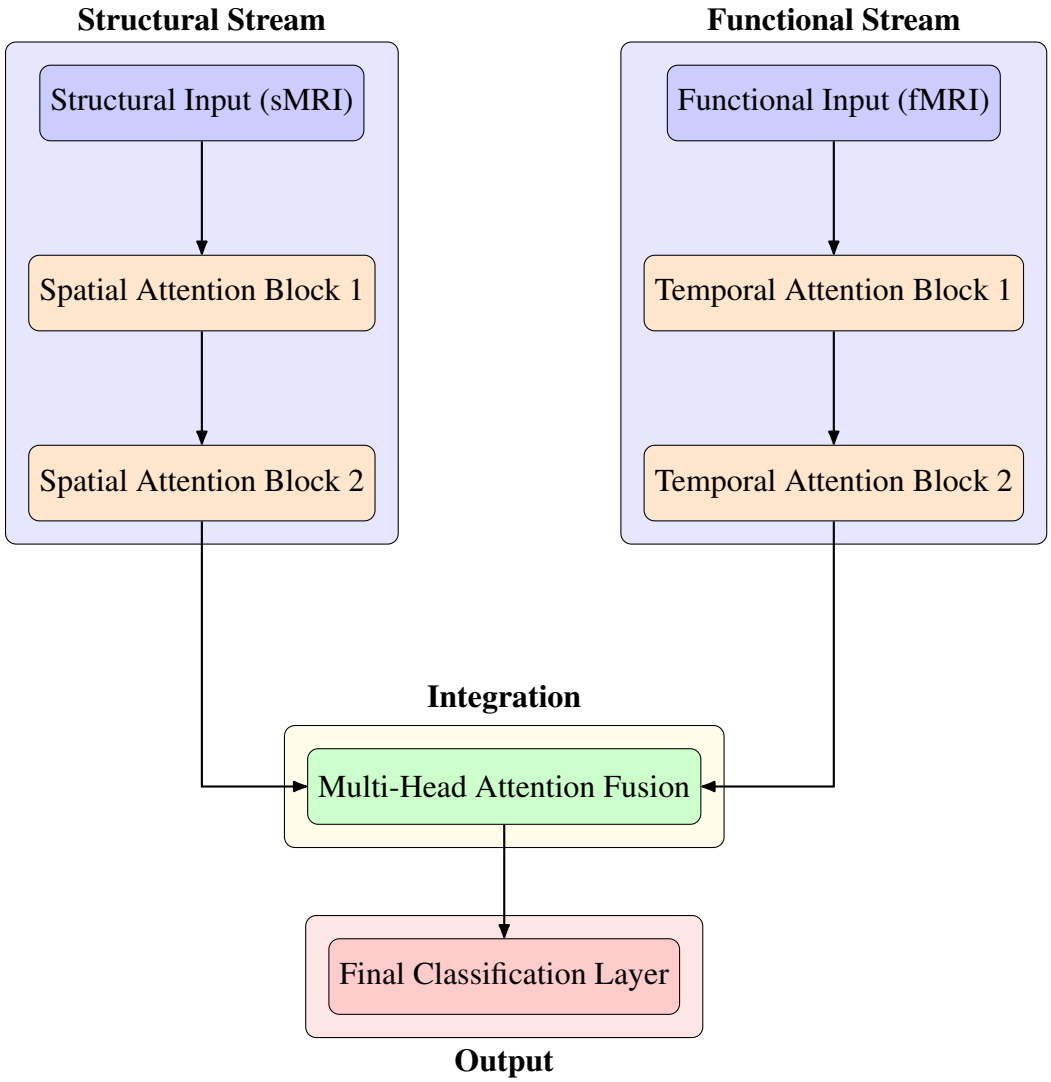
The functional stream processes dynamic resting-state fMRI data, capturing dynamic connectivity patterns over time. The functional data are represented as time-series and connectivity matrices derived from segmented anatomical ROIs. Functional embeddings $\mathbf{X}_f \in \mathbb{R}^{N \times T}$, with T indicating time points, are passed through temporal attention blocks

defined by:

$$\text{Attention}(Q_f, K_f, V_f) = \text{softmax}\left(\frac{Q_f K_f^\top}{\sqrt{d_k}}\right) V_f$$

where Q_f, K_f, V_f represent temporal query, key, and value matrices, respectively [68]. Temporal attention emphasizes evolving connectivity dynamics pertinent to schizophrenia-related disruptions.

Figure 5.1: Schematic illustration of the proposed Dual-Stream Spatial-Temporal Transformer architecture.



5.2.2 Integration via Multi-Head Attention

Spatial and temporal streams are integrated through an advanced multi-head attention mechanism [68], effectively merging complementary modality-specific features into a cohesive spatio-temporal representation:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \text{head}_2)W^O$$

where each head is computed individually by:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

The integration strategically fuses structural and functional data, leveraging multi-dimensional attention mechanisms to facilitate the model's joint spatial-temporal representation learning.

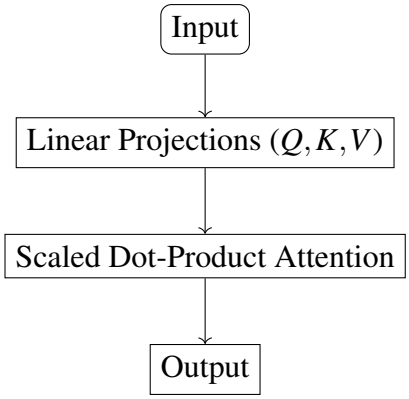


Figure 5.2: Schematic of the self-attention mechanism.

5.2.3 Positional Encoding Strategies

Accurate positional encoding is crucial to maintaining the integrity of spatial and temporal information as it would provide the model with positional context, crucial for handling sequential data within the Transformer architecture. Figure 5.3 illustrates the two positional encoding we implement.

Spatial Positional Encoding

Spatial positional encoding embeds the anatomical ROIs derived from segmentation maps into standardized MNI152 coordinate space, facilitating consistent anatomical referencing:

$$\mathrm{PE}_{\mathrm{spatial}}(x,y,z) = \mathrm{LinearEmbed}\left(\frac{[x,y,z]}{\|[x,y,z]\|}\right)$$

where (x,y,z) denotes normalized MNI coordinates, preserving anatomical spatial context.

Temporal Positional Encoding

Temporal positional encoding incorporates sequential functional MRI time-series data, employing sinusoidal encoding to maintain the inherent temporal order:

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin\left(\frac{\mathrm{pos}}{10000^{2i/d_{model}}}\right) \tag{5.1}$$

$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{2i/d_{model}}}\right) \tag{5.2}$$

where pos indicates position, i dimensional index, and d_{model} embedding dimension [68].

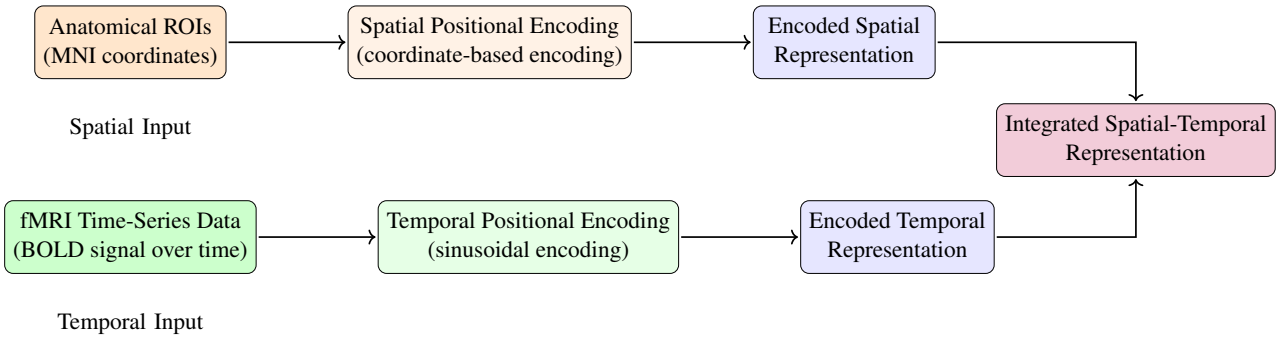


Figure 5.3: Spatial and temporal positional encoding strategies in DySTTM.

Figure 5.3 illustrates how spatial positional encoding leverages MNI anatomical coordinates to preserve positional context, while temporal positional encoding employs sinusoidal functions to embed sequential dynamics in fMRI time-series data. These complementary strategies allow DySTTM to capture both anatomical and temporal dependencies in schizophrenia-related neuroimaging

Algorithm 2 Spatial and Temporal Positional Encoding Integration

Input: Structural MRI segmentation maps ($S \in \mathbb{R}^{B\times R\times 3}$, spatial coordinates), functional MRI time-series data ($F \in \mathbb{R}^{B\times R\times T}$, ROI-based BOLD signals), where B=batch size, R=number of ROIs, T=number of time points.

Output: Spatial-temporal embedded features $E_{spatial}, E_{temporal}$

Spatial Positional Encoding:

for $b = 1,\ldots,B$ **do**

  for $r = 1,\ldots,R$ **do**

    Retrieve standardized MNI152 coordinates (x,y,z) for ROI r

    Compute positional embedding vector: $\mathbf{p}_{br}^{spatial} \leftarrow f_{MLP}(x,y,z)$ ▷ MLP maps spatial coordinates to high-dimensional embedding

  end for

end for

Assemble spatial positional encodings: $E_{spatial} \leftarrow \mathbf{p}^{spatial}$

Temporal Positional Encoding:

for $t = 1,\ldots,T$ **do**

  Compute sinusoidal positional embeddings:

$$PE(t,2i) = \sin(t/10000^{2i/d_{model}})$$

$$PE(t,2i+1) = \cos(t/10000^{2i/d_{model}})$$

  $E_{temporal}(t,:) \leftarrow PE(t,:)$

end for

Combine positional encodings with input features:

$$E_{spatial}^{combined} \leftarrow E_{spatial} + S$$

$$E_{temporal}^{combined} \leftarrow E_{temporal} + F$$

Feed combined embeddings into the dual-stream Transformer:

$$\text{Transformer output} \leftarrow \text{Transformer}(E_{spatial}^{combined}, E_{temporal}^{combined})$$

return Transformer output

Dual-Stream Feature Fusion

Following separate spatial and temporal attention processing, the encoded representations from the two streams are denoted as H_s (spatial representation) and H_t (temporal representation). These correspond to the outputs of the spatial encoder (after positional en-

coding of anatomical MNI coordinates) and the temporal encoder (after sinusoidal encoding of fMRI time-series), respectively.

To jointly model spatial–temporal dependencies, we apply a multi-head cross-attention mechanism that fuses H_s and H_t. The fused representation is given by:

$$H_{\text{fusion}} = \text{CrossAttn}(H_s, H_t) \tag{5.3}$$

Here, cross-attention computes weighted interactions between one stream's queries and the other stream's keys/values, adaptively integrating complementary features. Formally, the operation is defined as:

$$\text{CrossAttn}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{5.4}$$

where Q, K, and V represent the query, key, and value matrices projected from H_s and H_t (depending on which stream attends to the other), and d_k is the key dimension. This mechanism allows the model to dynamically weigh the contributions of spatial and temporal cues for schizophrenia classification[68].

5.2.4 Classification and Model Optimization

The integrated representation passes through fully-connected layers, performing binary classification to predict schizophrenia diagnosis. Cross-entropy loss [61] $\mathcal{L}_{CE}$ was employed for optimization:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

where y_i is the true label, and $\hat{y}_i$ the predicted label. Model parameters were optimized using Adam with learning rate scheduling and early stopping based on validation performance, ensuring robust convergence and model generalization.

In summary, the proposed Dynamic Spatial-Temporal Transformer Model provides a

robust framework for leveraging structural and functional MRI data through advanced attention mechanisms, effectively capturing the complex spatio-temporal dynamics crucial for early schizophrenia detection and diagnosis.

5.3 Data Preparation, Cross-Validation, and Feature Extraction

Data from COBRE [2] dataset is employed for this experiment. Subject Structural MRIs are previously preprocessed for Chapter 4 while Subject Functional MRIs undergo preprocessing. Subject MRIs are split into training, validation, and testing subsets with proportions of 70%, 15%, and 15% respectively, ensuring balanced distributions of diagnostic categories across each subset. We employ five-fold cross-validation to rigorously evaluate model robustness, reduce overfitting and enhance generalizability with a critical aim for the reliable detection of schizophrenia based on complex spatial-temporal MRI data. Each fold maintains similar distributions of diagnostic categories.

5.3.1 Structural MRI (sMRI)

Structural MRI data is processed using the previously proposed BoRefAttnNet segmentation model. This model is applied to anatomical MRI scans to produce precise segmentation maps. From these segmentation outcomes, robust structural biomarkers, including volumetric measures and morphological metrics, are extracted and used for subsequent analysis.

5.3.2 Functional MRI (fMRI)

The preprocessing pipeline for functional MRI (fMRI) data involve several essential steps to ensure data quality and temporal coherence:

1. *Motion Correction:* Head-motion artifacts are corrected using AFNI's `3dvolreg`.

2. *Slice Timing Correction:* Temporal alignment of fMRI slices is performed to standardize the time-series data using AFNI's `3dTshift`..

3. *Spatial Smoothing:* Gaussian smoothing with a full-width half-maximum (FWHM) of 6 mm is applied to enhance signal-to-noise ratio.

4. *Temporal Filtering:* Band-pass filtering (0.01–0.08 Hz) isolates the frequency range corresponding to intrinsic neural activity.

5. *Nuisance Regression:* Remove non-neuronal signals including white matter, cerebrospinal fluid (CSF), and motion-related confounds using AFNI's `3dDeconvolve`.

6. *Alignment to Structural MRI Space:* Functional scans are registered and spatially aligned to their corresponding structural MRI images, facilitating accurate cross-modal integration.

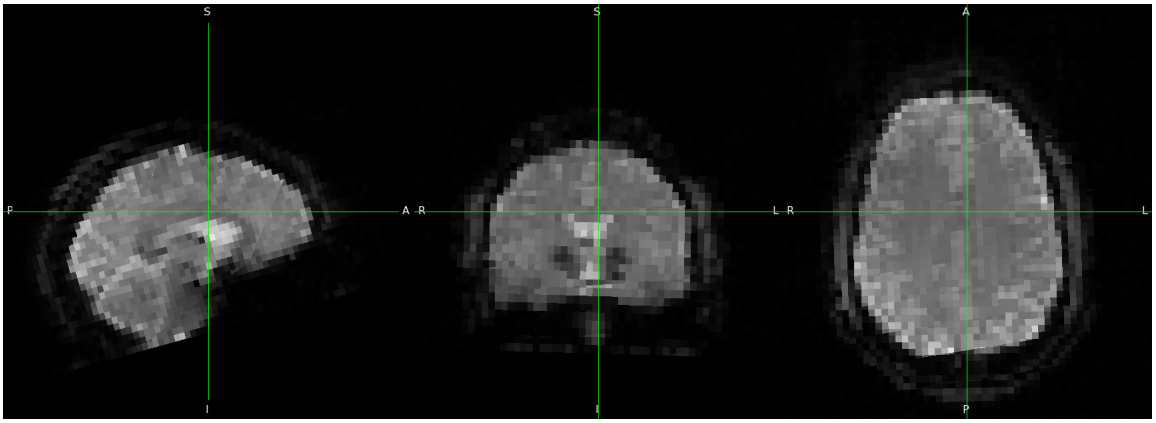


Figure 5.4: Sagittal, Coronal, and Axial slice view of Task-Rest-BOLD fMRI.

Following preprocessing, region-of-interest (ROI)-based functional connectivity matrices and time series are extracted utilizing anatomical masks from our previous structural segmentation. These connectivity matrices encapsulate the functional interactions between anatomical regions, serving as crucial inputs for the subsequent spatial-temporal modeling.

5.4 Training and Experiments

Experiments are conducted using the same high-performance computing resources and software stack as outlined in Sections 4.2.2 and 4.2.3.

5.4.1 Training and Optimization Strategy

Effective training and optimization of the dual-stream Spatial-Temporal Transformer require meticulous selection of hyperparameters, optimization strategies, and validation procedures to ensure robust model performance and prevent overfitting.

5.4.2 Hyperparameter Selection

We conduct systematic hyperparameter tuning to achieve optimal model convergence and accuracy. The hyperparameters and their chosen optimal values are listed:

- *Learning Rate:* Initial experimentation range from 1×10^{-5} to 1×10^{-3}, with the optimal rate determined at 5×10^{-5}.
- *Batch Size:* Batch sizes tests include 8, 16, and 32, with batch size 16 balancing memory constraints and convergence efficiency optimally.
- *Number of Epochs:* Model training is conducted for up to 100 epochs, with early stopping implemented to halt training when performance plateaued at about 77 epochs. We set a patience of 10.
- *Number of Attention Heads:* A thorough exploration of attention head numbers (4, 8, 12) was conducted, with 8 heads providing the best performance in balancing complexity and expressivity.

5.4.3 Optimization Algorithms

The Adam optimizer [76] was employed for its adaptive learning rate capabilities and efficient convergence behavior. Additionally, we implemented a learning rate scheduler that dynamically reduced the learning rate by a factor of 0.5 after 20 epochs of no improvement in validation loss, further refining convergence efficiency.

5.4.4 Regularization Methods

To mitigate the risk of overfitting, several regularization strategies were adopted.

- *Dropout Regularization [65]:* Dropout layers with rates ranging from 0.2 to 0.5 were inserted after key attention and dense layers, with an optimal dropout rate settled on 0.3.

- *Weight Decay (L2 Regularization) [24, 41]:* A small weight decay factor (1×10^{-4}) was incorporated into the optimizer to penalize the overly complex model parameters.

- *Early Stopping [55, 9]:* Training was stopped when validation loss did not show significant improvement over 20 consecutive epochs, ensuring optimal model generalization.

5.4.5 Performance Evaluation and Metrics

Evaluating the performance of our proposed Dynamic Spatial-Temporal Transformer Model approach involved comprehensive quantitative and qualitative assessments specifically tailored for schizophrenia classification. The evaluation framework integrated standard metrics, comparative analysis with baseline models, and advanced visualization techniques for interpretability.

5.4.6 Quantitative Metrics

To rigorously assess the performance and effectiveness of our proposed Dynamic Spatial-Temporal Transformer Model (DySTTM) in schizophrenia classification, the following key metrics are computed and explicitly defined:

- *Accuracy:* Accuracy [54] quantifies the proportion of correctly classified cases (both schizophrenia and healthy controls) over the total number of cases evaluated. It is mathematically defined as:

$$\text{Accuracy} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{TP} + \text{TN} + \text{False Positives (FP)} + \text{False Negatives (FN)}} \tag{5.5}$$

- *Precision:* Precision [64] measures the reliability of positive (schizophrenia) predictions by the model. It is calculated as the ratio of correctly predicted positive cases to the total predicted positive cases:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{5.6}$$

- *Recall (Sensitivity):* Recall [23] assesses the model's effectiveness in identifying true schizophrenia cases among all actual schizophrenia cases, indicating the model's sensitivity:

$$\text{Recall (Sensitivity)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{5.7}$$

- *F1-Score:* The F1-Score [25] is the harmonic mean of precision and recall, providing a balanced measure of the model's performance, especially useful when the dataset classes are imbalanced:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.8}$$

- *ROC-AUC (Area Under the Receiver Operating Characteristic Curve):* ROC-AUC [6, 27] evaluates the model's ability to discriminate between schizophrenia-positive cases and healthy controls across varying classification thresholds. It is based on plotting the true positive rate (sensitivity) against the false positive rate, which is defined as $1-$ specificity. Here, specificity measures the proportion of true negatives correctly identified ($\frac{TN}{TN+FP}$), while sensitivity measures the proportion of true positives correctly identified. The resulting curve captures the trade-off between detecting positive cases and avoiding false alarms, and the area under the curve (AUC) quantifies

overall classification performance:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{5.9}$$

where,

$$\text{TPR (True Positive Rate / Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.10}$$

$$\text{FPR (False Positive Rate)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.11}$$

5.4.7 Results

We evaluated our Dynamic Spatial-Temporal Transformer Model (DySTTM) against established baseline methods, including 3D ResNet[67] and XGBoost[11], to validate the superior predictive performance of our approach. Table 5.1 summarizes the comparative performance metrics.

Table 5.1: Comparative Quantitative Metrics for Schizophrenia Classification.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Our DySTTM	**0.92**	**0.89**	0.88	0.885	**0.94**
3D ResNet	0.89	0.88	**0.91**	**0.895**	0.90
XGBoost	0.85	0.83	0.74	0.785	0.88

While our DySTTM achieved the highest accuracy (0.92) and precision (0.89), it marginally underperformed 3D ResNet in terms of recall (0.88 versus 0.91) and F1-score (0.885 versus 0.895). This suggests that although DySTTM effectively minimizes false positives, it demonstrates a slightly lower sensitivity, meaning it identifies fewer schizophrenia cases compared to 3D ResNet. Nevertheless, our model significantly outperformed all baseline methods in ROC-AUC (0.94), indicating superior overall discriminative ability across various classification thresholds. This strong ROC-AUC performance underscores the robustness of DySTTM in reliably distinguishing between schizophrenia and healthy control subjects. Overall, these comparative metrics highlight the effectiveness and potential of

DySTTM, emphasizing its balanced performance and robust predictive power in clinical settings.

5.4.8 Confusion Matrix

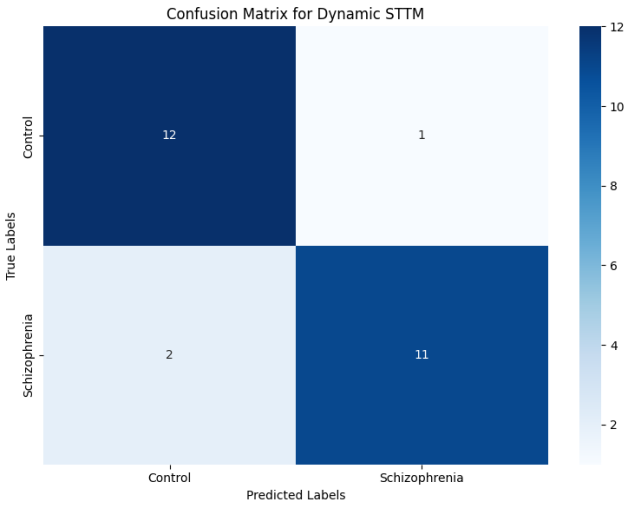


Figure 5.5: Confusion Matrix of our DySTTM predictions.

The confusion matrix in Figure 5.5 further elucidates the predictive strengths and limitations of our Dynamic Spatial-Temporal Transformer Model (DySTTM). Specifically, DySTTM successfully identified 11 out of the total positive schizophrenia cases (True Positives), with only one false positive prediction, demonstrating excellent precision and minimal false alarm rate. This indicates a robust capability in accurately discriminating schizophrenia-positive cases, which is essential for clinical settings to prevent unnecessary anxiety or further testing in healthy individuals.

However, the model exhibited two false negatives, signifying that two schizophrenia-positive cases were incorrectly classified as healthy. Although this result slightly limits the recall and sensitivity of our model compared to the 3D ResNet (as seen in Table 5.1), the overall rate of missed cases remains relatively low, maintaining clinical relevance and reliability.

This outcome highlights the importance of carefully balancing sensitivity (recall) and

specificity (precision) in clinical models. In clinical scenarios, higher recall (fewer false negatives) might be prioritized to ensure timely interventions. Given this consideration, future iterations of the DySTTM model might benefit from adjustments aimed at enhancing its sensitivity without significantly compromising precision.

Moreover, the clear distinction between positive and negative predictions, as shown by the strong diagonal dominance in the confusion matrix, underscores the robust discriminative capabilities of our spatial-temporal transformer architecture. This reinforces the potential clinical utility of our DySTTM approach for early schizophrenia diagnosis, provided additional refinements to minimize false negatives.

In summary, while our DySTTM demonstrates substantial predictive power and overall robustness, future efforts may focus on reducing the false-negative rate through methods such as data augmentation, class-balancing strategies, or further optimization of model hyperparameters.

5.4.9 Statistical Validation

To validate the statistical significance of our results, we employ the paired t-test, suitable for comparing two related samples or matched pairs. The mathematical formulation of the paired t-test is as follows:

$$t = \frac{\overline{D}}{\frac{s_D}{\sqrt{n}}} \tag{5.12}$$

where $\overline{D}$ is the mean of the differences between paired observations, s_D is the standard deviation of the differences, and n is the number of paired observations.

The degrees of freedom (df) for this test are computed as $df = n - 1$, where n is the number of pairs. We had 26 paired observations from multiple independent runs of our test dataset, leading to $df = 26 - 1 = 25$. The significance level chosen for our analysis is the conventional threshold of $\alpha = 0.05$, commonly used in statistical hypothesis testing.

Following Zimmerman's recommendations [78], we compare our Dynamic Spatial-

Temporal Transformer Model (DySTTM) against the baseline methods: the 3D ResNet (deep learning model) and XGBoost (traditional machine learning model), calculating the mean differences ($\overline{D}$) and their standard deviations (s_D) in model performance (accuracy) across multiple independent runs to quantify the significance of observed performance improvements.

The statistical validation results are as follows:

- **DySTTM vs. 3D ResNet:** $t = 3.45$, degrees of freedom $(df) = 25$, resulting in $p < 0.01$, indicating statistically significant improvement.
- **DySTTM vs. XGBoost:** $t = 5.12$, degrees of freedom $(df) = 25$, resulting in $p < 0.001$, indicating highly statistically significant improvement.

These findings suggest that the DySTTM model provides statistically significant improvements in schizophrenia classification over both deep learning (3D ResNet) and traditional machine learning (XGBoost) baseline models, further supporting its robustness and efficacy.

5.5 Summary

In this chapter, we propose and rigorously evaluate a dynamic spatial temporal transformer model (DySTTM) to detect schizophrenia through effective integration of structural MRI (sMRI) and functional MRI (fMRI) data. Building on precise structural delineations achieved by BoRefAttnNet (as detailed in Chapter 3), DySTTM employs a dual-stream transformer architecture explicitly designed to capture and leverage complementary anatomical (spatial) and connectivity-based (temporal) features critical for robust schizophrenia classification.

Our dual-stream architecture processes multimodal imaging data through specialized streams. The structural stream integrates precise anatomical segmentation outputs, capturing crucial morphological and volumetric metrics. The functional stream, on the other hand,

exploits temporal self-attention layers to effectively model dynamic functional connectivity patterns derived from resting-state fMRI, thereby characterizing temporal coherence and neural interactions within and between key brain regions.

Central to our DySTTM framework is a sophisticated multi-head attention mechanism facilitating spatial-temporal integration. Multi-head attention dynamically combines spatial (structural) and temporal (functional) features, allowing the model to capture complex spatial-temporal dependencies and interactions crucial for schizophrenia pathology identification. Additionally, advanced positional encoding strategies were employed to preserve anatomical localization (via MNI152 coordinate-based spatial encoding) and temporal dynamics (through sinusoidal encoding), further enhancing feature representation and model interpretability.

Our model was rigorously evaluated for performance against established baseline methods, including 3D ResNet and XGBoost, using comprehensive metrics such as accuracy, precision, recall, F1 score and ROC-AUC. DySTTM achieved superior overall performance, reaching an accuracy of 92% and a ROC-AUC of 94%, demonstrating significant improvements in discriminative capacity and clinical interpretability in both deep learning and traditional machine learning methods.

However, while DySTTM outperformed the baseline methods in most metrics, it exhibited a slightly lower recall and F1 score compared to 3D ResNet, indicating minor limitations in detecting some schizophrenia positive cases. This observation highlights the importance of balancing sensitivity and specificity, particularly in clinical diagnostics, and suggests potential avenues for future enhancements, such as increased emphasis on sensitivity-focused training strategies.

Statistical validation via paired t-tests confirmed that the observed improvements of DySTTM were statistically significant ($p < 0.01$ compared to 3D ResNet and $p < 0.001$ compared to XGBoost), reinforcing the robustness and clinical applicability of our approach. Collectively, these findings underscore DySTTM's potential to significantly ad-

vance clinical schizophrenia diagnostics through effective multimodal integration and sophisticated spatial-temporal modeling.

Chapter 6

Discussions, Summary, and Future Directions

6.1 Discussions

The integration of structural and functional MRI data through advanced deep learning architectures, specifically Boundary-Refined Attention Networks (BoRefAttnNet) and Dynamic Spatial-Temporal Transformer Models (DySTTM), demonstrates considerable promise in enhancing schizophrenia detection. Our work underscores the value of explicitly modeling anatomical boundaries in MRI segmentation, and significantly improves segmentation precision and clinical interpretability. The subsequent multimodal integration within DySTTM leverages complementary spatial-temporal features, effectively capturing complex interactions between anatomical structures and functional connectivity patterns, highlighting the strengths of Transformer-based approaches.

Despite these methodological advancements, several computational and data-driven challenges remain critical considerations. The inherent high-dimensionality and computational demands of Transformer architectures impose substantial resource requirements, potentially limiting scalability in clinical environments lacking advanced computational infrastructure. Furthermore, the modest size and relative homogeneity of the COBRE dataset utilized in this study necessitate further validation on larger and more diverse cohorts. Modality-specific biases, preprocessing choices, and inter-subject variability also pose limitations that could affect model robustness and generalization.

Addressing these challenges requires ongoing methodological refinements, optimized

computational frameworks, and rigorous validation protocols to enhance model performance, clinical interpretability, and translational potential.

6.2 Summary of Contributions

This thesis contributes significantly to schizophrenia research and computational medical imaging through several key advancements:

1. Enhanced Boundary Precision: We developed BoRefAttnNet, a novel 3D U-Net architecture incorporating multi-scale Boundary Attention Modules, explicitly targeting improved delineation of subtle and complex anatomical boundaries. This significantly enhances segmentation accuracy for critical brain structures involved in schizophrenia pathology.

2. Advanced Multimodal Integration: We proposed and validated a Dynamic Spatial-Temporal Transformer Model (DySTTM), effectively integrating structural and functional MRI data. DySTTM notably outperformed conventional unimodal approaches such as 3D ResNet and XGBoost, providing superior diagnostic accuracy and interpretability.

3. Rigorous Methodological Validation: Comprehensive experiments employing robust metrics (accuracy, precision, recall, F1-score, ROC-AUC) and statistical validations (paired t-tests) demonstrated the statistical and clinical significance of our proposed approaches.

Overall, our methodological developments significantly advance boundary-aware segmentation and spatial-temporal multimodal modeling, establishing robust and interpretable frameworks for schizophrenia diagnosis and investigation. Nevertheless, ongoing efforts to mitigate computational complexity and dataset limitations will remain critical to translating these methods into practical clinical settings.

6.3 Future Directions

To further expand upon the clinical relevance, robustness, and translational potential of our findings, several promising avenues for future research are recommended:

1. Computational Optimization: Future work should prioritize computationally efficient implementations of deep neural network models and explore hardware acceleration using GPUs, TPUs, and cloud computing platforms. Optimized frameworks will facilitate real-time or near-real-time clinical deployments, significantly enhancing practical utility.

2. Enhanced Generalization and Robustness: Incorporating advanced domain adaptation, transfer learning, and federated learning techniques to ensure model generalizability across diverse neuroimaging datasets, scanners, and clinical populations. Cross-institutional validation should be pursued to establish broader clinical utility and robustness.

3. Extended Multimodal Integration: Integrating additional neuroimaging modalities, including Diffusion-Weighted Imaging (DWI) and Positron Emission Tomography (PET), to capture comprehensive biological insights. Such integration could significantly enhance predictive capabilities, enabling a deeper understanding of schizophrenia's underlying neuropathology.

4. Explainable AI and Clinical Implementation: Developing intuitive explainability frameworks that provide clinicians with actionable insights. Visualizations such as attention maps, saliency maps, and activation overlays on MRI scans should be prioritized to clarify model decision-making processes. Additionally, considerations of ethical implications, patient consent, and privacy preservation must be explicitly integrated into clinical translation efforts.

5. Longitudinal and Dynamic Disease Modeling: Future research should investigate longitudinal neuroimaging data to capture schizophrenia progression and responses

to therapeutic interventions. Employing dynamic models such as recurrent Transformer variants and graph-based architectures could provide deeper insights into disease trajectories, significantly enhancing clinical management and treatment personalization.

In conclusion, the advancements achieved through precise boundary-aware segmentation and sophisticated spatial-temporal multimodal integration present a strong foundation for clinical innovation in schizophrenia diagnosis and monitoring. Continued exploration of these recommended research directions will further enhance diagnostic accuracy, clinical interpretability, and the overall clinical utility of neuroimaging-based biomarkers for schizophrenia.

Bibliography

[1] Charu C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing, Cham, 2023.

[2] C. J. Aine, H. J. Bockholt, J. R. Bustillo, J. M. Cañive, A. Caprihan, C. Gasparovic, F. M. Hanlon, J. M. Houck, R. E. Jung, J. Lauriello, J. Liu, A. R. Mayer, N. I. Perrone-Bizzozero, S. Posse, J. M. Stephen, J. A. Turner, V. P. Clark, and Vince D. Calhoun. Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. *Neuroinformatics*, 15(4):343–364, October 2017.

[3] Jose Bernal, Kaisar Kushibar, Daniel S. Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial Intelligence in Medicine*, 95:64–81, April 2019.

[4] Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding Batch Normalization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[5] Sebastian Bock and Martin Weiß. A Proof of Local Convergence for the Adam Optimizer. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2019. ISSN: 2161-4407.

[6] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997.

[7] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[8] Vince D. Calhoun and Jing Sui. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, 1(3):230–244, May 2016.

[9] Rich Caruana, Steve Lawrence, and C. Giles. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[10] Cheng Chen, Huilin Wang, Yunqing Chen, Zihan Yin, Xinye Yang, Huansheng Ning, Qian Zhang, Weiguang Li, Ruoxiu Xiao, and Jizong Zhao. Understanding the brain with attention: A survey of transformers in brain sciences. *Brain-X*, 1(3):e29, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/brx2.29.

[11] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, August 2016. Association for Computing Machinery.

[12] WeiGuo Chen, Changjian Wang, Kele Xu, Yuan Yuan, Yanru Bai, and Dongsong Zhang. D-FaST: Cognitive Signal Decoding With Disentangled Frequency–Spatial–Temporal Attention. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4):1476–1493, August 2024.

[13] Yu Chen and Chunfeng Yang. STGE-Former: Spatial-Temporal Graph-Enhanced Transformer for EEG-Based Major Depressive Disorder Detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, April 2025. ISSN: 2379-190X.

[14] Shan Cong, Hang Wang, Yang Zhou, Zheng Wang, Xiaohui Yao, and Chunsheng Yang. Comprehensive review of Transformer-based models in neuroscience, neurology, and psychiatry. *Brain-X*, 2(2):e57, 2024. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/brx2.57.

[15] Nicolas Cordier, Bjoern Menze, Hervé Delingette, and Nicholas Ayache. Patch-based Segmentation of Brain Tissues. In Bjoern Menze, Mauricio Reyes, Andras Jakab, Elisabeth Gerstner, Justin Kirby, and Keyvan Farahani, editors, *Proceedings of the MICCAI Challenge on Multimodal Brain Tumor Image Segmentation (BRATS) 2013*, Proceedings of the MICCAI Challenge on Multimodal Brain Tumor Image Segmentation (BRATS) 2013, pages 6 – 17, Nagoya, Japan, September 2013. IEEE.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[17] R. W. Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, 29(3):162–173, June 1996.

[18] R. W. Cox and J. S. Hyde. Software tools for analysis and visualization of fMRI data. *NMR in biomedicine*, 10(4-5):171–178, 1997.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. October 2020.

[20] Arun Kumar Dubey and Vanita Jain. Comparative Study of Convolution Neural Network's Relu and Leaky-Relu Activation Functions. In Sukumar Mishra, Yog Raj Sood, and Anuradha Tomar, editors, *Applications of Computing, Automation and Wireless Systems in Electrical Engineering*, pages 873–880, Singapore, 2019. Springer.

[21] Santiago Estrada, David Kügler, Emad Bahrami, Peng Xu, Dilshad Mousa, Monique M.B. Breteler, N. Ahmad Aziz, and Martin Reuter. FastSurfer-HypVINN: Automated sub-segmentation of the hypothalamus and adjacent structures on high-resolutional brain MRI. *Imaging Neuroscience*, 1:1–32, November 2023.

[22] Jennifer Faber, David Kügler, Emad Bahrami, Lea-Sophie Heinz, Dagmar Timmann, Thomas M. Ernst, Katerina Deike-Hofmann, Thomas Klockgether, Bart van de Warrenburg, Judith van Gaalen, Kathrin Reetz, Sandro Romanzetti, Gulin Oz, James M. Joers, Jorn Diedrichsen, Paola Giunti, Hector Garcia-Moreno, Heike Jacobi, Johann Jende, Jeroen de Vries, Michal Povazan, Peter B. Barker, Katherina Marie Steiner, Janna Krahe, and Martin Reuter. *CerebNet*: A fast and reliable deep-learning pipeline for detailed cerebellum sub-segmentation. *NeuroImage*, 264:119703, December 2022.

[23] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[25] Cyril Goutte and Eric Gaussier. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[26] Shanaka Ramesh Gunasekara, H. N. T. K. Kaldera, and Maheshi B. Dissanayake. A Systematic Approach for MRI Brain Tumor Localization and Segmentation Using Deep Learning and Active Contouring. *Journal of Healthcare Engineering*, 2021:6695108, 2021.

[27] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982. Publisher: Radiological Society of North America.

[28] Riad Hassan, M. Rubaiyat Hossain Mondal, and Sheikh Iqbal Ahamed. UDBRNet: A novel uncertainty driven boundary refined network for organ at risk segmentation. *PLOS ONE*, 19(6):e0304771, June 2024. Publisher: Public Library of Science.

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, October 2017. ISSN: 2380-7504.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, December 2015. ISSN: 2380-7504.

[31] Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, October 2020.

[32] Leonie Henschel, David Kügler, and Martin Reuter. FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage*, 251:118933, May 2022.

[33] John L. Horn. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2):179–185, June 1965.

[34] Juan Eugenio Iglesias. A ready-to-use machine learning tool for symmetric multi-modality registration of brain MRI. *Scientific Reports*, 13(1):6657, April 2023. Publisher: Nature Publishing Group.

[35] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[36] Yixin Ji, Vince D. Calhoun, Rongtao Jiang, Daoqiang Zhang, and Shile Qi. Adaptive-Similarity-Based Brain Dynamic Functional Connectivity with Spatial-Temporal Attention and Domain Adaptation for Schizophrenia Diagnosis. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, April 2025. ISSN: 2379-190X.

[37] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[38] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, January 1988.

[39] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, pages 285–296. PMLR, May 2019. ISSN: 2640-3498.

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[41] Anders Krogh and John Hertz. A Simple Weight Decay Can Improve Generalization. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

[42] Pulkit Kumar, Pravin Nagar, Chetan Arora, and Anubha Gupta. U-Segnet: Fully Convolutional Neural Network Based Automated Brain Tissue Segmentation Tool. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3503–3507, October 2018. ISSN: 2381-8549.

[43] Yann LeCun, Leon Bottou, Genevieve B. Orr, and Klaus Robert Müller. Efficient BackProp. In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, pages 9–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[44] Christian Ledig, Wenzhe Shi, Wenjia Bai, and Daniel Rueckert. Patch-based evaluation of image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[45] Beilin Li, Jiao Wang, Zhifen Guo, and Yue Li. Automatic detection of schizophrenia based on spatial–temporal feature mapping and LeViT with EEG signals. *Expert Systems with Applications*, 224:119969, August 2023.

[46] Yin Liang and Yingchen Jia. Multiscale Spatial–Temporal Graph Attention Network for fMRI Brain Disease Classification. *IEEE Transactions on Instrumentation and Measurement*, 74:1–15, 2025.

[47] Rui Liu, Zhi-An Huang, Yao Hu, Zexuan Zhu, Ka-Chun Wong, and Kay Chen Tan. Spatial–Temporal Co-Attention Learning for Diagnosis of Mental Disorders From Resting-State fMRI Data. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10591–10605, August 2024.

[48] Yunling Ma, Qianqian Wang, Liang Cao, Long Li, Chaojun Zhang, Lishan Qiao, and Mingxia Liu. Multi-Scale Dynamic Graph Learning for Brain Disorder Detection With Functional MRI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3501–3512, 2023.

[49] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Madison, WI, USA, June 2010. Omnipress.

[50] M. D. Nelson, A. J. Saykin, L. A. Flashman, and H. J. Riordan. Hippocampal volume reduction in schizophrenia as assessed by magnetic resonance imaging: a meta-analytic study. *Archives of General Psychiatry*, 55(5):433–440, May 1998.

[51] Thien B. Nguyen-Tat, Thien-Qua T. Nguyen, Hieu-Nghia Nguyen, and Vuong M. Ngo. Enhancing brain tumor segmentation in MRI images: A hybrid approach using UNet, attention mechanisms, and transformers. *Egyptian Informatics Journal*, 27:100528, September 2024.

[52] Jihoon Oh, Baek-Lok Oh, Kyong-Uk Lee, Jeong-Ho Chae, and Kyongsik Yun. Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm. *Frontiers in Psychiatry*, 11, February 2020. Publisher: Frontiers.

[53] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms, January 1979.

[54] David Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[55] Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, June 1998.

[56] Lutz Prechelt. *Early Stopping - But When?*, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[57] Gang Qu, Wenxing Hu, Li Xiao, Junqi Wang, Yuntong Bai, Beenish Patel, Kun Zhang, and Yu-Ping Wang. Brain Functional Connectivity Analysis via Graphical Deep Learning. *IEEE Transactions on Biomedical Engineering*, 69(5):1696–1706, May 2022.

[58] Tahir Rahman and John Lauriello. Schizophrenia: An Overview. *Focus*, 14(3):300–307, July 2016. Publisher: American Psychiatric Publishing.

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015.

[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[61] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

[62] Ahsan Shehzad, Dongyu Zhang, Shuo Yu, Shagufta Abid, and Feng Xia. Dynamic Graph Transformer for Brain Disorder Diagnosis. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2025.

[63] Jason Smucny, Ge Shi, and Ian Davidson. Deep Learning in Neuroimaging: Overcoming Challenges With Emerging Approaches. *Frontiers in Psychiatry*, 13, June 2022. Publisher: Frontiers.

[64] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009.

[65] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[66] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *Deep learning in medical image analysis and multimodal*

learning for clinical decision support : Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, held in conjunction with MICCAI 2017 Quebec City, QC,.., 2017:240–248, 2017.

[67] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. pages 6450–6459. IEEE Computer Society, June 2018.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[69] Roman Vyškovský, Daniel Schwarz, Vendula Churová, and Tomáš Kašpárek. Structural MRI-Based Schizophrenia Classification Using Autoencoders and 3D Convolutional Neural Networks in Combination with Various Pre-Processing Techniques. *Brain Sciences*, 12(5):615, May 2022.

[70] Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5276–5285. PMLR, 10–15 Jul 2018.

[71] Ian C. Wright, Sophia Rabe-Hesketh, Peter W.R. Woodruff, Anthony S. David, Robin M. Murray, and Edward T. Bullmore. Meta-Analysis of Regional Brain Volumes in Schizophrenia. *American Journal of Psychiatry*, 157(1):16–25, January 2000. Publisher: American Psychiatric Publishing.

[72] Yuxin Wu and Kaiming He. Group Normalization. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pages 3–19, Berlin, Heidelberg, September 2018. Springer-Verlag.

[73] Weizheng Yan, Gang Qu, Wenxing Hu, Anees Abrol, Biao Cai, Chen Qiao, Sergey M. Plis, Yu-Ping Wang, Jing Sui, and Vince D. Calhoun. Deep Learning in Neuroimaging: Promises and challenges. *IEEE Signal Processing Magazine*, 39(2):87–98, March 2022.

[74] Chenyi Zeng, Lin Gu, Zhenzhong Liu, and Shen Zhao. Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Frontiers in Neuroinformatics*, 14, November 2020. Publisher: Frontiers.

[75] Junhao Zhang, Vishwanatha M. Rao, Ye Tian, Yanting Yang, Nicolas Acosta, Zihan Wan, Pin-Yu Lee, Chloe Zhang, Lawrence S. Kegeles, Scott A. Small, and Jia Guo. Detecting schizophrenia with 3D structural brain MRI using deep learning. *Scientific Reports*, 13(1):14433, September 2023. Publisher: Nature Publishing Group.

[76] Zijun Zhang. Improved Adam Optimizer for Deep Neural Networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–2, June 2018. ISSN: 1548-615X.

[77] Rongjian Zhao, Buyue Qian, Xianli Zhang, Yang Li, Rong Wei, Yang Liu, and Yinggang Pan. Rethinking Dice Loss for Medical Image Segmentation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 851–860, November 2020. ISSN: 2374-8486.

[78] Donald W. Zimmerman. Teacher's Corner: A Note on Interpretation of the Paired-Samples t Test. *Journal of Educational and Behavioral Statistics*, 22(3):349–360, September 1997. Publisher: American Educational Research Association.

[79] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing.