ScholarWorks@GSU

Attentive Multi-modal Learning: Unifying Visual and Clinical Insights for Medical Image Classification and Report Generation

Authors	Nagur Shareef Shaik
Citation	Nagur Shareef Shaik (2025). "Attentive Multi-modal Learning: Unifying Visual and Clinical Insights for Medical Image Classification and Report Generation." Thesis, Georgia State University. https://doi.org/10.57709/tw0z-py92
DOI	https://doi.org/10.57709/tw0z-py92
Download date	2026-11-04 12:20:58
Link to Item	https://hdl.handle.net/20.500.14694/15456

Attentive Multi-modal Learning: Unifying Visual and Clinical Insights for Medical Image Classification and Report Generation

by

Nagur Shareef Shaik

Under the Direction of Dong Hye Ye, Ph.D.

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

Master of Science

in the College of Arts and Sciences

Georgia State University

2025

ABSTRACT

This work investigates the integration of diverse multi-modal data-spanning medical imaging, clinical text, and genetic information-through scalable attention networks and vision-language models to improve diagnostic accuracy and automate medical report generation. The core hypothesis is that attentive multi-modal learning can seamlessly unify visual and clinical insights, thereby enhancing the precision of disease diagnosis and facilitating efficient medical report generation. Focusing on neuroimaging and retinal imaging, this research introduces novel architectures and methodologies that leverage deep learning techniques to synthesize and analyze multi-modal data. We propose an innovative attention mechanism called *Spatial Sequence Attention* (SSA) to identify subtle brain morphological changes in structural MRIs (sMRIs) that are otherwise imperceptible through manual analysis, specifically targeting the diagnosis of complex cognitive impairments like schizophrenia. Building on this, we extend the approach to integrate additional multi-modal data, incorporating Functional Network Connectivity (FNC) data and Single Nucleotide Polymorphisms (SNPs) genomic information. To achieve this, we introduce the *Multi-modal Imaging Genomics Transformer* (MIGTrans), which leverages attention-based mechanisms to uncover structural brain abnormalities, functional connectivity disruptions, and relevant genetic markers associated with schizophrenia, enhancing classification accuracy and interpretability. Finally, we extend this multi-modal framework to retinal imaging, introducing the Multi-Modal Medical Transformer (M3T) to generate clinically relevant medical reports. M3T efficiently integrates retinal images with diagnostic text, improving both the quality and coherence of the generated reports.

INDEX WORDS: Multi-modal Learning, Attention Networks, Vision-Language Models (VLM), Schizophrenia, Medical Report Generation.

Copyright by
Nagur Shareef Shaik
2025

Attentive Multi-modal Learning: Unifying Visual and Clinical Insights for Medical Image Classification and Report Generation

by

Nagur Shareef Shaik

Committee Chair: Dong Hye Ye

Committee: Vince D Calhoun

Hemanth D Venkateswara

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2025

DEDICATION

I wholeheartedly dedicate this thesis to my parents, **Haseena Shaik**, **Lal Bude Shaik**, and my sister, **Reshma Shaik**, who have been the pillars of strength and support throughout my life. Their relentless sacrifices, enduring love, and unwavering belief in me have not only made my education possible but have also instilled in me the resilience and curiosity to explore the uncharted territories of deep learning and computational sciences. They gave what little they had to offer me a wealth of opportunity, and it is their spirit that has fueled my academic journey. I also dedicate this work to my advisor, **Dr. Dong Hye Ye**, whose guidance, encouragement, and astute insights have been invaluable in shaping both this research project and my broader academic career. Dr. Ye has been more than just a supervisor to me; he has been a mentor in the truest sense of the word. His willingness to assist in any way possible, and his commitment to academic excellence have continually pushed me to reach greater heights. I am incredibly fortunate to have had such a dedicated and insightful advisor to guide me through this intricate and challenging field. Finally, I dedicate this thesis to my fellow researcher and friend, **Teja Krishna Cherukuri**, whose collaboration, innovative ideas, and unwavering support have been instrumental in overcoming challenges and completing this work. This thesis is a culmination of the lessons learned, the challenges overcome, and the countless hours of research and experimentation. It is a testament to the enduring love and sacrifices of my family, the insightful guidance of Dr. Ye, and the invaluable collaboration of Teja. To them, I dedicate this body of work.

ACKNOWLEDGMENTS

I extend my deepest gratitude to all who supported and guided me throughout this academic journey. First and foremost, I thank my advisor, **Dr. Dong Hye Ye**, for his invaluable guidance, expertise, and unwavering support, which shaped this research and my growth as a scholar. I am also immensely grateful to **Dr. Vince Calhoun**, Director of TReNDS, for providing an inspiring research environment and broadening my perspective on medical imaging and genomics. I acknowledge **Dr. Hemanth Venkateswara** for his foundational teachings during my master's program, which were pivotal in my understanding of deep learning. A special thanks to my best friend and fellow researcher, **Teja Krishna Cherukuri**, for his collaboration, innovative ideas, and constant support. He has been an endless source of encouragement and constructive criticism which were invaluable in overcoming challenges.

I am deeply grateful to *my parents* and *sister* for their unconditional love and encouragement, which kept me motivated. I also thank my bachelor's advisor, **Dr. Jyostna Devi Bodapati**, for her continued mentorship and guidance. To all my friends - especially **Nagendra Satya Chowdary Vattikuti**, and my dear *roommates* - thank you for your companionship and laughter, which kept me grounded. Finally, I thank my *relatives* and *fellow researchers* for their constant support. This thesis is a culmination of the collective support of everyone mentioned above. I am thankful to each of you for making this scholarly work a reality.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . . **v**

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

1 INTRODUCTION **1**

1.1 Schizophrenia Classification 2

1.2 Medical Report Generation 5

1.3 Hypothesis 6

1.3.1 Objectives 7

1.3.2 Impact 7

1.4 Research Contributions 8

1.5 Thesis Outline 8

2 RELATED WORK **10**

2.1 Evolution of Uni-modal Approaches for Schizophrenia Classification 10

2.2 Advancements in Multi-modal Integration for Schizophrenia Diagnosis 11

2.3 Deep Learning based Medical Report Generation Models 13

2.3.1 Multi-modal Learning for Medical Report Generation 14

2.4 Challenges and Research Gaps 14

2.4.1 Challenges in Schizophrenia Classification 15

2.4.2 Challenges in Automated Medical Report Generation 16

3 SPATIAL SEQUENCE ATTENTION NETWORK **17**

3.1 Morphological Representation Learning 18

3.2 Spatial Sequence Attention 19

3.3 Regularized Feed Forward Neural Network 21
3.4 Experiments & Results 22
3.4.1 Dataset 22
3.4.2 Implementation Details 22
3.4.3 Quantitative Evaluation 23
3.4.4 Qualitative Evaluation 24
4 MULTI-MODAL IMAGING GENOMICS TRANSFORMER 26
4.1 Genomic Connectome Representation Learning 26
4.1.1 Genomic Encoder 26
4.1.2 Connectome Encoder 27
4.2 Morphological Representation Learning 28
4.2.1 Structural MRI Encoder 28
4.2.2 Spatial Sequence Attention 29
4.3 Fusion Transformer 30
4.3.1 Genomic Connectome TransFusor 30
4.3.2 Genomic Connectome sMRI TransFusor 31
4.4 Experiments & Results 33
4.4.1 Dataset and Pre-processing 33
4.4.2 Implementation Details 33
4.4.3 Quantitative Evaluation 34
4.4.4 Qualitative Evaluation 35
5 MULTI-MODAL MEDICAL TRANSFORMER 37
5.1 Visual Encoder 38
5.1.1 Convolutional Base 38
5.1.2 Lesion Contextual Gate 38
5.2 Keywords Encoder 40
5.3 TransFusion Encoder 42

5.4 Medical Description Decoder 43
5.5 Experiments & Results 45
5.5.1 Dataset 45
5.5.2 Pre-processing 46
5.5.3 Implementation Details 46
5.5.4 Quantitative Evaluation 47
5.5.5 Qualitative Evaluation 48
6 CONCLUSION **51**
6.1 Future Research Directions 51
REFERENCES **53**
APPENDIX **59**
A Papers Published 59

LIST OF TABLES

Table 3.1 Comparison of Five-fold Cross-validation Metrics (in %) for SZ Classification 23

Table 4.1 Classification performance comparison of proposed model with baselines 34

Table 5.1 Results demonstrating the effect of incorporating Keywords as input to the model and integrating Lesion Contextual Gate with Visual features, along with Context-aware Attention to Keyword Embeddings; 47

Table 5.2 Comparative Study of Recent Best Models with Proposed Context Gating Vision Transformer trained on DeepEyeNet Dataset, highlighting the impact of incorporating keyword embeddings and Attention on Visual features; 48

LIST OF FIGURES

Figure 1.1 Illustration of morphological changes in hippocampal and thalamic volumes. The figure includes four structural MRI (sMRI) scans: hippocampal atrophy with reduced hippocampal volume, normal hippocampal volume, normal thalamic volume, and reduced thalamic volume, gray matter density in the context of schizophrenia. 3

Figure 3.1 The proposed Spatial Sequence Attention Network (SSANet) architecture extracts morphological features from 3D T1-weighted MRI images (121x145x121 voxels) using a pre-trained 3D DenseNet121 model. The Spatial Sequence Attention (SSA) module then captures and accentuates complex spatial interactions and relationships within the brain volumes. Finally, a Feed Forward Neural Network (FFNN) classifies the images as schizophrenia or healthy. 17

Figure 3.2 Visualization of (A) Attention Heatmaps; (B) Gray Matter Density Maps of a normal subject; 24

Figure 3.3 Visualization of (A) Attention Heatmaps; (B) Gray Matter Density Maps of a SZ subject; SSANet correctly gives attention around the disease impacted areas such as frontal lobe and temporal lobe. Note the decreased gray matter density in the SZ subject compared with the normal subject in those regions. 25

Figure 4.1 Architecture of Multi-modal Imaging Genomics Transformer (MIGTrans) - employing **Genomic & Connectome Encoders** for genomic & connectome representation learning; **Structural MRI encoder** and **Spatial Sequence Attention** for morphological representation learning; **Fusion Transformer** for multi-modal step-wise attentive integration 27

Figure 4.2 **Attention maps of sMRI** highlighting the frontal and temporal lobes in SZ samples, contrasting with thalamic regions in HC samples; **Functional brain networks** highlighting important connections in default mode (DMN) and cognitive control (CON) specific to SZ and HC compared with other subcortical (SCN), auditory (ADN), sensorimotor (SMN), visual (VSN), and cerebellum (CBN) networks. 35

Figure 5.1 Architecture of proposed Multi-Modal Medical Transformer (M3T); **Visual Encoder** – learns attention-based representations from retinal images; **Keyword Encoder** – learns clinical-context embeddings from diagnostic keywords; **TransFusion Encoder** – integrates visual attention features and clinical-context embeddings, leveraging both visual and semantic information; **Medical Description Generation Decoder** – generates coherent and meaningful medical descriptions by attending to relevant visual and semantic cues, ensuring contextually appropriate and diagnostically relevant outputs; . 37

Figure 5.2 Visualization of Lesion Contextual Gate Attention heatmaps and corresponding overlays, highlighting the visual insights from color fundus, gray fundus, OCT, and Auto-Fluorescent retinal scan images 49

Figure 5.3 Comparison of Actual, Non-Local Attention, and M3T medical descriptions 50

CHAPTER 1
INTRODUCTION

The human brain and eyes, two of the most complex and vital organs, are at the forefront of modern medical research, offering profound insights into both localized and systemic health. The retina, often considered an extension of the brain, can reveal chronic conditions like cardiovascular risks, diabetes, and neurodegenerative diseases through detailed scans. Similarly, the brain's structural and functional changes can provide early indicators of systemic disorders such as hypertension and metabolic syndromes. Studying these organs, not only uncover their own pathologies but also gain critical insights into the health of other vital systems, enabling a more comprehensive approach to diagnosis and treatment.

Recent advancements in medical imaging and artificial intelligence (AI) have revolutionized our ability to study these organs with unprecedented precision, transforming the landscape of healthcare diagnostics Shaik et al. (2024b) Shaik et al. (2025b). In neuroimaging, cutting-edge modalities such as structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), and diffusion tensor imaging (DTI) now enable the detection of subtle structural and functional changes in the brain, providing new avenues for diagnosing and treating complex conditions like schizophrenia (SZ). These technologies reveal intricate details about brain morphology, connectivity, and activity, offering unprecedented insights into neurological disorders. Similarly, in ophthalmology, retinal imaging techniques such as fundus photography, optical coherence tomography (OCT), and fluorescein angiography have emerged as critical tools for the early detection and management of

vision-threatening diseases, such as diabetic retinopathy and age-related macular degeneration, if left untreated, can lead to irreversible blindness. These imaging modalities provide high-resolution views of the retina, enabling clinicians to identify pathological changes at their earliest stages and intervene effectively.

These innovations, powered by AI, have not only enhanced diagnostic accuracy but also introduced automated systems capable of translating complex imaging data into actionable clinical insights. However, the growing complexity of medical data—spanning imaging, genomics, and clinical text—presents both a challenge and an opportunity Thapaliya et al. (2025a) Thapaliya et al. (2025b). By integrating multi-modal data, we can transcend the limitations of single-modality approaches, uncovering deeper insights into disease mechanisms and revolutionizing the way we diagnose, treat, and understand human health. This thesis explores the frontier of attentive multi-modal learning, aiming to bridge the gap between visual data and clinical context to advance precision medicine for schizophrenia classification and medical report generation Cherukuri et al. (2025) Goldblatt et al. (2024b) Goldblatt et al. (2024a).

1.1 Schizophrenia Classification

Schizophrenia (SZ) is a severe and chronic mental disorder affecting approximately 24 million people worldwide, as reported by the World Health Organization (WHO) (Cannon et al. 2000). Characterized by persistent hallucinations, cognitive impairments, and disruptive behavior, SZ significantly impacts the daily lives of individuals and poses a substantial

burden on healthcare systems. Diagnosing SZ is particularly challenging due to the overlap of its symptoms with other psychiatric disorders, necessitating advanced tools for accurate and early detection (Park et al. 2023) Sapkota et al. (2024) Sapkota et al. (2025).

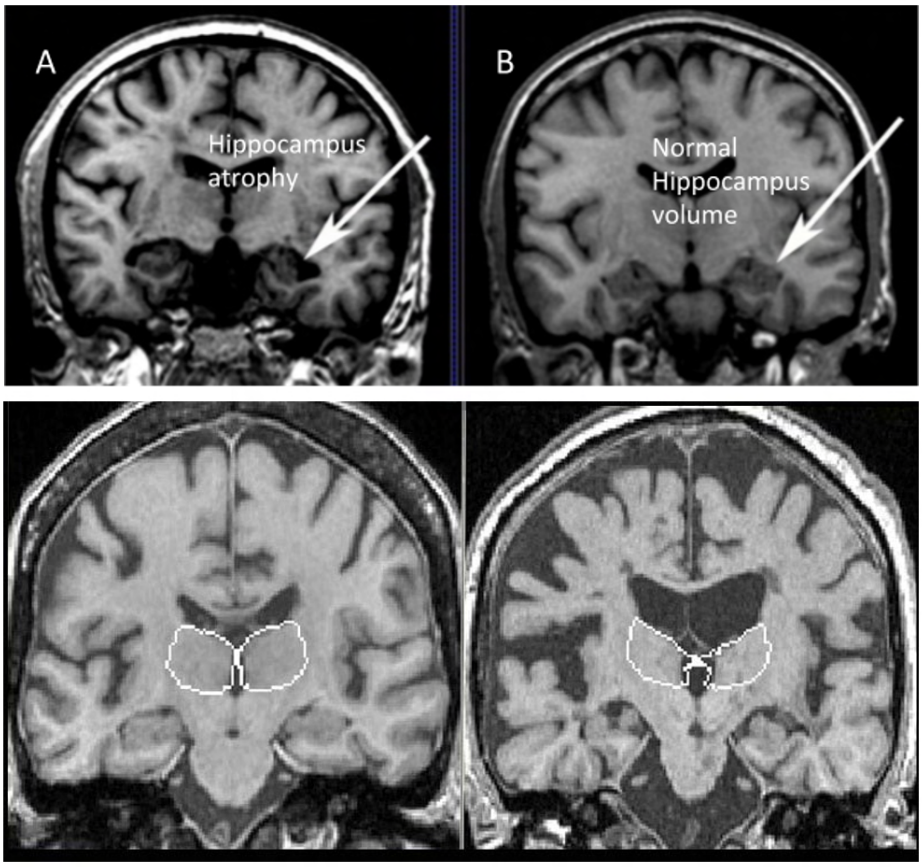


Figure 1.1 Illustration of morphological changes in hippocampal and thalamic volumes. The figure includes four structural MRI (sMRI) scans: hippocampal atrophy with reduced hippocampal volume, normal hippocampal volume, normal thalamic volume, and reduced thalamic volume, gray matter density in the context of schizophrenia.

Structural magnetic resonance imaging (sMRI) has been instrumental in identifying consistent brain abnormalities in SZ patients, such as reduced hippocampal and thalamic volumes, increased globus pallidus volumes, and widespread reductions in gray matter density (Haukvik et al. 2013). Figure 1.1, illustrates morphological changes in hippocampal and

thalamic volumes. These structural differences are key indicators of brain abnormalities associated with complex disorders like schizophrenia. However, these structural changes alone provide limited insights into the disorder, highlighting the need for more comprehensive diagnostic approaches. Complementing sMRI with functional connectome data and genomic markers offers a promising avenue for unraveling the intricate pathophysiology of SZ.

Functional MRI (fMRI) enables the construction of brain networks, such as Functional Network Connectivity (FNC), which captures the dynamic interactions between different brain regions (Chen et al. 2023). In SZ patients, disruptions in these functional networks—such as altered connectivity in the default mode network (DMN) and frontoparietal networks—have been linked to cognitive deficits and symptom severity (Hung et al. 2024). These functional connectivity changes provide a deeper understanding of how brain regions communicate, offering insights into the neural mechanisms underlying SZ that structural imaging alone cannot reveal.

Genomic markers, particularly Single Nucleotide Polymorphisms (SNPs), further enhance diagnostic precision by identifying genetic variations associated with SZ (Kim & Misra 2007). These markers can reveal heritable risk factors and molecular pathways contributing to the disorder, bridging the gap between genetic predisposition and observed brain abnormalities. For example, certain SNPs have been linked to disruptions in synaptic plasticity and neurotransmitter systems, which are critical for brain function (Kanyal et al. 2023).

By integrating sMRI, FNC, and SNPs, clinicians can gain a more holistic view of SZ, capturing not only structural and functional brain alterations but also the genetic underpinnings

of the disorder. This multi-modal approach enables the identification of biomarkers that reflect the complex interplay between brain structure, functional connectivity, and genetic risk factors, providing a deeper understanding of SZ pathophysiology.

1.2 Medical Report Generation

The global prevalence of vision-threatening retinal diseases, such as Diabetic Retinopathy, Diabetic Macular Edema, and Age-related Macular Degeneration, has risen significantly, with over 500 million people projected to be affected by 2040 (World Health Organization 2023). These conditions, often leading to irreversible blindness, are caused by complications like sub-retinal neo-vascularization and retinal detachment (Shaik & Cherukuri 2021). Traditional diagnostic methods rely on resource-intensive processes, expensive equipment, and the expertise of ophthalmologists, creating barriers to timely and accurate medical reporting. This highlights the urgent need for automated solutions that can streamline diagnosis, improve accessibility to care (Zheng et al. 2012).

Automated medical report generation from retinal images has emerged as a promising approach to address these challenges. By leveraging advancements in Computer Vision and Natural Language Processing (NLP), this task aims to generate precise and clinically relevant descriptions of retinal pathologies. Early efforts in this domain have drawn inspiration from image captioning techniques, where Convolutional Neural Networks (CNNs) extract visual features from images, and Recurrent Neural Networks (RNNs) generate textual descriptions (Xu et al. 2015; Kamal et al. 2020). However, despite significant progress in natural image

captioning using advanced models like Transformers, applying these techniques to retinal images remains challenging due to the complexity of retinal pathology and the scarcity of annotated medical data (Beddiar et al. 2022). Recent developments in multi-modal learning have shown promise in improving caption generation by integrating visual features with diagnostic keywords or textual context (Herdade et al. 2019). For medical report generation, combining retinal images with diagnostic text can enhance the semantic relevance of the generated descriptions. However, efficiently learning the interplay between visual context and clinical semantics remains a critical challenge.

1.3 Hypothesis

This research is founded on the hypothesis that attentive multi-modal learning frameworks, utilizing state-of-the-art deep learning techniques, can effectively integrate visual insights from medical imaging (such as structural, functional MRI and retinal scans) with clinical context from genomic biomarkers textual data. By unifying these diverse modalities, the proposed frameworks aim to significantly improve diagnostic accuracy for complex disorders like schizophrenia and enhance the quality of automated medical report generation for retinal images. This approach seeks to capture the intricate relationships between brain structure, functional connectivity, genetic variations, and clinical context, enabling a more comprehensive understanding of disease pathophysiology and advancing precise, interpretable, and efficient medical diagnostics.

1.3.1 Objectives

This hypothesis is operationalized through the following three objectives:

1. *Develop advanced attention mechanisms* to process high-dimensional medical imaging data, such as structural MRIs (sMRIs), with a focus on identifying subtle morphological changes and emphasizing critical spatial interactions for improved sensitivity.

2. *Integrate multi-modal data*, including structural MRI (sMRI), functional MRI (fMRI)-based Functional Network Connectivity (FNC), and Single Nucleotide Polymorphisms (SNPs) from genomic data, to uncover structural, functional, and genetic markers associated with complex disorders like schizophrenia (SZ).

3. *Automate medical report generation* by combining retinal imaging data with diagnostic text to produce precise, coherent, and clinically relevant medical descriptions, enhancing efficiency and accuracy in ophthalmology diagnostics.

1.3.2 Impact

This research has the potential to transform medical diagnostics by developing scalable, interpretable, and efficient tools for integrating multi-modal data. By unifying visual insights with clinical context, the proposed frameworks advance precision medicine, enabling earlier and more accurate diagnosis of complex disorders like schizophrenia and improving automated medical reporting. These innovations promise to enhance patient outcomes, streamline healthcare workflows, and provide actionable insights, ultimately driving a paradigm shift toward personalized, data-driven healthcare solutions.

1.4 Research Contributions

1. Developed the Spatial Sequence Attention Network (SSANet) to process sMRI data, identifying subtle brain morphological changes in schizophrenia (SZ) through a novel Spatial Sequence Attention (SSA) mechanism. SSANet outperforms existing methods, enhancing sensitivity to SZ-related abnormalities.

2. Introduced the Multi-modal Imaging Genomics Transformer (MIGTrans), integrating sMRI, fMRI-based FNC, and genomic SNPs to uncover structural, functional, and genetic markers of SZ.

3. Proposed the Multi-Modal Medical Transformer (M3T) for retinal imaging, combining visual features with diagnostic text to generate precise, clinically relevant medical reports, meeting ophthalmologists' standards for accuracy and coherence.

1.5 Thesis Outline

This thesis is structured to systematically address the challenges and opportunities in multi-modal medical diagnostics and automated reporting. Chapter 2 reviews existing literature on schizophrenia (SZ) classification, genomic integration, and automated medical reporting, providing a foundation for the proposed methodologies. Chapter 3 introduces the Spatial Sequence Attention (SSA) mechanism, detailing its architecture and evaluating its performance on SZ-specific structural MRI (sMRI) dataset. Chapter 4 presents the Multi-modal Imaging Genomics Transformer (MIGTrans), focusing on its stepwise integration of sMRI, functional

MRI (fMRI), and genomic data, along with its interpretability and clinical insights. Chapter 5 extends the proposed methodologies to retinal imaging, introducing the Multi-Modal Medical Transformer (M3T) for automated medical report generation and highlighting its advancements in precision and coherence. Finally, Chapter 6 summarizes the key findings, discusses their clinical implications, and outlines future research directions.

CHAPTER 2
RELATED WORK

The field of medical diagnostics has witnessed significant advancements in recent years, driven by the integration of medical imaging, artificial intelligence (AI), and multi-modal data analysis. This chapter reviews existing literature on schizophrenia (SZ) classification, multi-modal data integration, and automated medical report generation, highlighting key trends, challenges, and limitations.

2.1 Evolution of Uni-modal Approaches for Schizophrenia Classification

Early research in SZ classification relied on machine learning techniques, such as support vector machines (SVMs), to differentiate SZ patients from healthy controls based on gray matter density (Schnack et al. 2014). Other studies employed extended Granger causality and Kendall's tau rank correlation coefficients for feature extraction and selection, followed by SVM-based classification (Wismüller & Vosoughi 2021; Wismüller et al. 2023). While these methods demonstrated promising results, they were limited by their reliance on handcrafted features and linear models, which struggled to capture the complex, non-linear relationships in brain data.

The advent of deep learning, particularly convolutional neural networks (CNNs) (LeCun et al. 1995), has revolutionized SZ classification by enabling automated feature extraction from 3D sMRI scans. For instance, Hu et al. utilized 3D CNNs and pre-trained 2D CNN models to detect SZ using sMRI data, achieving significant improvements in classification accuracy (Hu et al. 2022). However, most existing methods rely on holistic features learned

by CNNs without leveraging attention mechanisms to emphasize critical regions, potentially limiting their performance.

Attention mechanisms, such as the Squeeze and Excitation Network (SENet), have been introduced to enhance feature extraction by dynamically recalibrating channel importance in feature maps (Hu et al. 2018). SENet has shown promise in medical image classification but faces limitations in adapting to higher-dimensional data and localizing critical regions effectively (Shaik & Cherukuri 2022a). To address these challenges, this study explores the design of novel attention mechanism, called Spatial Sequence Attention Network (SSANet), which can selectively emphasizes SZ-specific information at spatial locations, improving sensitivity to morphological alterations across brain volume.

2.2 Advancements in Multi-modal Integration for Schizophrenia Diagnosis

To gain a deeper understanding of SZ, researchers have increasingly turned to multi-modal data integration, combining structural MRI (sMRI), functional MRI (fMRI), and genomic biomarkers. Functional MRI (fMRI) enables the construction of brain networks, such as Functional Network Connectivity (FNC), which captures the dynamic interactions between different brain regions (Chen et al. 2023). In SZ patients, disruptions in these functional networks—such as altered connectivity in the default mode network (DMN) and frontoparietal networks—have been linked to cognitive deficits and symptom severity (Hung et al. 2024).

Genomic markers, particularly Single Nucleotide Polymorphisms (SNPs), further enhance diagnostic precision by identifying genetic variations associated with SZ (Kim & Misra 2007).

These markers can reveal heritable risk factors and molecular pathways contributing to the disorder, bridging the gap between genetic predisposition and observed brain abnormalities. For example, certain SNPs have been linked to disruptions in synaptic plasticity and neurotransmitter systems, which are critical for brain function (Kanyal et al. 2023).

Existing studies have explored various approaches to multi-modal data integration, including simple concatenation of features (Jafri et al. 2008), attentional feature fusion (Dai et al. 2021), and advanced methods like Sparse Canonical Correlation Analysis (SCCA) (Du et al. 2018) and Graph Neural Networks (GNNs) (Chan et al. 2022). While these methods have shown promise, they often overlook complex inter-modality relationships or fail to capture non-linear interactions effectively. For instance, SCCA and GNNs focus on linear projections or graph-based representations, which may not fully exploit the complementary information across modalities (Zhou et al. 2023).

Recent advancements in vision transformers (ViTs) have provided a non-linear approach to learning attentive representations, enabling the integration of functional and structural MRI features (Bi et al. 2023). However, integrating imaging data with genomic biomarkers remains a significant challenge due to their differing dimensionality, scale, and intricate relationships (Chuang et al. 2023). This reserach proposes Multi-modal Imaging Genomics Transformer (MIGTrans) to addresses these limitations by employing a step-wise attentive integration approach, strategically combining genomics, connectome, and sMRI features to enhance SZ diagnostic accuracy and interpretability.

2.3 Deep Learning based Medical Report Generation Models

The task of automated medical report generation is closely aligned with image captioning, a well-studied problem in computer vision and natural language processing (NLP). Early approaches employed Convolutional Neural Networks (CNNs) to extract visual features from images and Recurrent Neural Networks (RNNs) to generate textual descriptions (Xu et al. 2015; Kamal et al. 2020). These methods demonstrated promising results in natural image captioning but faced significant challenges when applied to medical images due to the complexity of retinal pathology and the scarcity of annotated medical data (Beddiar et al. 2022).

To address the scarcity of labeled data, researchers have turned to transfer learning, leveraging models pre-trained on large-scale natural image datasets like ImageNet and fine-tuning them for medical image analysis (Shaik & Cherukuri 2021, 2022a). This approach has proven effective in enhancing feature learning, particularly for medical image classification (Shaik & Cherukuri 2022b). Additionally, semi-supervised and self-supervised learning techniques have been explored to leverage unlabeled data, improving representation learning for downstream tasks (Misra & Maaten 2020; Li et al. 2023). Despite these advancements, generating clinically accurate and meaningful captions for retinal images remains a formidable challenge due to the intricate nature of retinal pathology and the need for domain-specific knowledge (Huang et al. 2021b).

2.3.1 Multi-modal Learning for Medical Report Generation

Recent developments in multi-modal learning have shown promise in improving caption generation by integrating visual features with diagnostic keywords or textual context (Herdade et al. 2019). For instance, the CLIP model, which aligns images and text in a shared embedding space, has been used to enhance the integration of visual and textual information (Radford et al. 2021). Similarly, contextualized keyword embeddings derived from techniques like GloVe and GPT2 have been integrated with retinal image features for caption generation, with GPT2 embeddings proving superior (Huang et al. 2021a). These methods aim to capture the semantic relationships between visual and textual data, improving the relevance and accuracy of generated reports.

However, existing approaches often struggle to efficiently learn the interplay between visual context and clinical semantics. For example, while non-local attention mechanisms have been used to highlight relevant regions in retinal images, they may not fully capture the nuanced relationships between visual features and diagnostic keywords (Huang et al. 2022). Similarly, Transformer-based models trained on visual features and keyword embeddings have demonstrated improved performance but still face challenges in generating coherent and clinically relevant descriptions (Wu et al. 2023).

2.4 Challenges and Research Gaps

Despite significant advancements in schizophrenia (SZ) classification and automated medical report generation, several challenges and research gaps remain unresolved in the literature.

These limitations highlight the need for innovative approaches to improve diagnostic accuracy, interpretability, and efficiency in medical diagnostics.

2.4.1 Challenges in Schizophrenia Classification

1. **Limited Sensitivity to Subtle Morphological Changes:** Existing methods, particularly those relying on holistic features from CNNs, often fail to emphasize critical regions in structural MRI (sMRI) data, limiting their sensitivity to subtle morphological changes associated with SZ.

2. **Ineffective Integration of Multi-modal Data:** While multi-modal approaches combining sMRI, fMRI, and genomic data have shown promise, many methods rely on simple concatenation or linear fusion techniques, which overlook complex inter-modality relationships and fail to capture non-linear interactions effectively.

3. **Lack of Interpretability:** Current models often lack interpretability, making it difficult for clinicians to understand the basis of diagnostic decisions, particularly when integrating genomic biomarkers with imaging data.

4. **Challenges in Genomic Data Integration:** Integrating imaging data with genomic biomarkers remains a significant challenge due to their differing dimensionality, scale, and intricate relationships, limiting the ability to uncover comprehensive biomarkers for SZ.

2.4.2 Challenges in Automated Medical Report Generation

1. **Limited Data:** The scarcity of annotated medical data, particularly for retinal image captioning, poses challenges in training robust deep learning models.

2. **Complex Pathology:** Retinal images often contain subtle and complex pathological features that are difficult to interpret, even for experienced clinicians, leading to inaccuracies in automated report generation.

3. **Inefficient Multi-modal Integration:** Existing methods often fail to efficiently integrate visual features with diagnostic text, limiting their ability to generate precise and coherent medical reports.

4. **Lack of Domain-Specific Training:** Many models rely on pre-trained features from natural images, which may not fully capture the unique characteristics of medical data, reducing the clinical relevance of generated reports.

This research addresses these challenges through the development of advanced attention mechanisms (SSANet), step-wise multi-modal integration (MIGTrans), and efficient multi-modal learning frameworks (M3T) for medical report generation. By prioritizing interpretability and scalability, the proposed frameworks aim to advance the state of the art in medical diagnostics, enabling more accurate, interpretable, and efficient tools for SZ classification and automated reporting. The following chapters delve into the methodologies and architectures of these frameworks, detailing their design, implementation, and evaluation.

CHAPTER 3

SPATIAL SEQUENCE ATTENTION NETWORK

This research introduces the Spatial Sequence Attention Network (SSANet), a deep learning architecture for automated SZ classification from 3D T1-weighted MRI images with limited training data. Figure 3.1 illustrates the SSANet architecture. In the subsequent sections, we provide detailed technical insights into each phase of the proposed methodology.

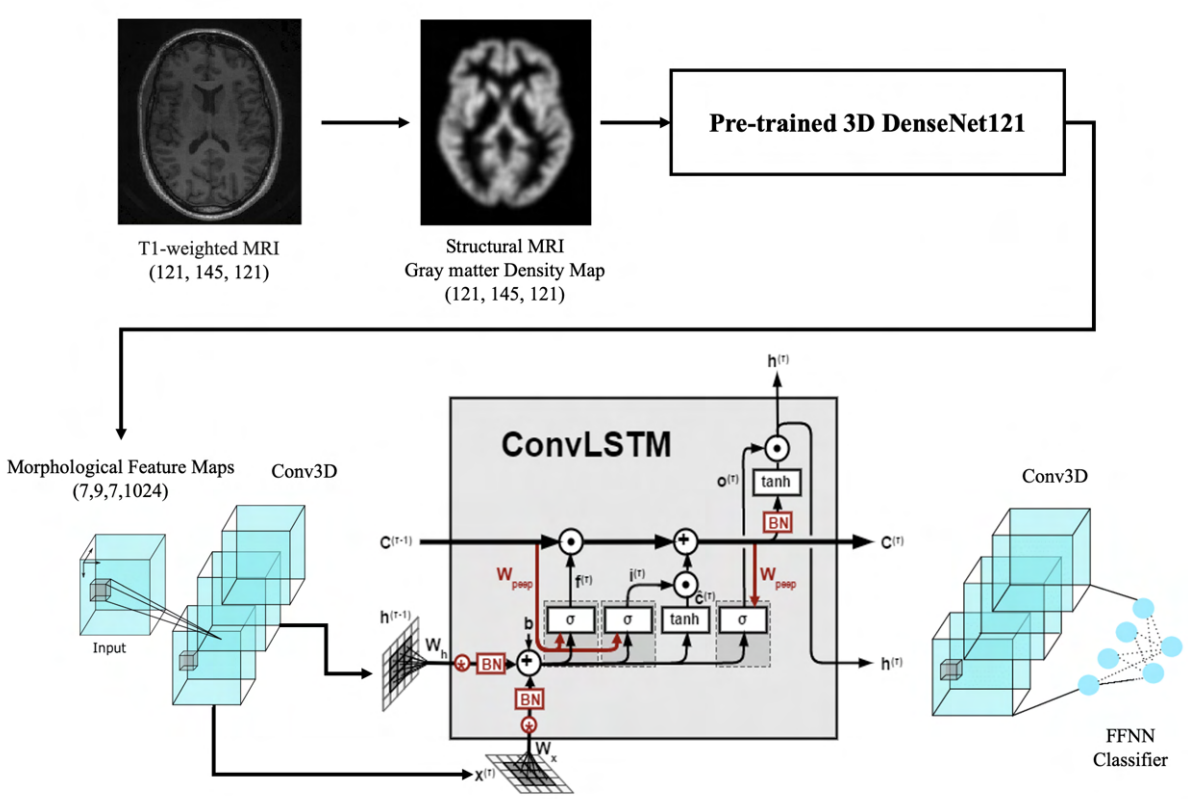


Figure 3.1 The proposed Spatial Sequence Attention Network (SSANet) architecture extracts morphological features from 3D T1-weighted MRI images (121x145x121 voxels) using a pre-trained 3D DenseNet121 model. The Spatial Sequence Attention (SSA) module then captures and accentuates complex spatial interactions and relationships within the brain volumes. Finally, a Feed Forward Neural Network (FFNN) classifies the images as schizophrenia or healthy.

3.1 Morphological Representation Learning

Over recent years, Convolutional Neural Networks (CNNs) have become prominent for learning representations from visual data. They employ a series of convolution operations to extract visual features by sliding a filter over the provided input image, capturing patterns and shapes through weighted summation of nearby pixel values, given the sufficient amount of training data (LeCun et al. 1995). The availability of training data is scarce in the medical domain. To address this, Transfer Learning was widely employed to extract representation from a limited dataset using pre-trained models. Following the paradigm, we leverage a pre-trained DenseNet, comprising 121 layers, to extract morphological features, augmenting its capabilities through fine-tuning specifically on gray matter density sMRIs.

In CNNs, the input from the previous layer $n-1$ is passed through a non-linear transformation denoted as H_n to generate the input for the current layer n, represented as $X^{(n)} = H_n(X^{(n-1)})$. This direct back-to-front gradient propagation can lead to the issue of vanishing gradients in the network. Addressing this, Huang et. al. presented the DenseNet with the dense block architecture. The n-th layer establishes connections with all preceding L layers, including the $n-1$-th layer, thereby facilitating improved gradient flow across the network (Huang et al. 2017). Most of the previous works focus on extracting penultimate fully connected layer features. Apart from this, we utilize the convolutional feature maps. These features retain spatial and volumetric information essential for discriminating SZ. This type of transfer learning method proves advantageous and adept at meeting the criteria for extracting morphological features (Bodapati et al. 2021).

$$X_i = DenseNet121(I_i) \quad \forall i \in \{1, 2, ...N\}. \tag{3.1}$$

From the equation 3.1, the input gray matter density image for subject i is denoted by I_i. It is a $(121 \times 145 \times 121)$ voxel image that is an input to a pre-trained 3D DenseNet121 model. The model first applies a 3D convolution with a $(7 \times 7 \times 7)$ kernel and pooling to the image. Then, it performs varying dense layers within each dense block. These layers include two 3D convolutions with different kernel sizes: $(1 \times 1 \times 1)$ and $(3 \times 3 \times 3)$, which reduce the feature map size and improve computation. Three transition layers (ReLU, batch normalization, and dimension reduction pooling) connect the dense blocks. This produces a $(7 \times 9 \times 7 \times 1024)$ output vector X_i for each subject i, which represents the morphological features of the subject's brain (Yu et al. 2022).

3.2 Spatial Sequence Attention

The Spatial Sequence Attention (Shaik & Cherukuri 2022a) is a special mechanism designed to capture spatial and volumetric dependencies in morphological feature maps X_i. It consists of three main components - a 3D convolutional (Conv3D) layer with 64 filters of a $(3 \times 3 \times 3)$ kernel, a ConvLSTM layer with 64 filters having $(3 \times 3 \times 3)$ contextual window, and a Conv3D layer with the same number of filters and kernel size as the input feature maps. Each component has its own significance in improving the provided feature maps. First Conv3D layer learns holistic features from the input. ConvolutionalLSTM is a variant of the LSTM (Long Short-Term Memory) with convolutional operations integrated into the

cell (SHI et al. 2015). It uses tanh and sigmoid activation functions to control the flow of information through the network and capture spatial information through sparse connections between the input and state transitions. This allows the model to learn intricate spatial interactions and relationships across volume in morphological feature maps. The significance of the current feature map is determined by considering both the current and past feature maps in the input volumetric sequence. The other Conv3D projects the output back to the original feature space.

High-level morphological feature representations obtained from the first Conv3D layer of SSA are processed through input gate (I_t), forget gate (F_t), and output gate (O_t), resulting in the evolution of the cell state ($\mathbf{C}_t$) and hidden state ($\mathbf{H}_t$) at a specific time step t. The future state of a specific feature map is determined by considering the inputs and previous states of neighboring feature maps within the receptive field. Convolutional operations in the transitions from input to state and from state to state facilitate this process. The ConvLSTM cell operates based on a set of fundamental equations involving convolution ($*$) and element-wise product ($\odot$). Resultants are attention-based feature representations that are subsequently passed as input to a regularized fully connected neural network to evaluate the presence of SZ in a given sMRI image.

$$I_t = \sigma(W_{xi} * \mathbf{X}_t + W_{hi} * \mathbf{H}_{t-1} + W_{ci} * \mathbf{C}_{t-1} + \mathbf{b}_i) \tag{3.2}$$

$$F_t = \sigma(W_{xf} * \mathbf{X}_t + W_{hf} * \mathbf{H}_{t-1} + W_{cf} * \mathbf{C}_{t-1} + \mathbf{b}_f) \tag{3.3}$$

$$\mathbf{C}_t = F_t \odot \mathbf{C}_t + I_t \odot \tanh(W_{xc} * \mathbf{X}_t + W_{hc} * \mathbf{H}_{t-1} + \mathbf{b}_c) \tag{3.4}$$

$$O_t = \sigma(W_{xo} * \mathbf{X}_t + W_{ho} * \mathbf{H}_{t-1} + W_{co} * \mathbf{C}_{t-1} + \mathbf{b}_o)\mathbf{H}_t \tag{3.5}$$

The above set of equations represents crucial steps in processing morphological feature maps while capturing spatial dependencies. The input gate (I_t) assesses the importance of each input feature map in sequence ($\mathbf{X}_t$), for inclusion in the cell state ($\mathbf{C}_t$). The forget gate (F_t) determines the extent to which features from the previous cell state ($\mathbf{C}_{t-1}$) should be retained. The cell state ($\mathbf{C}_t$) is then updated by selectively integrating features from the input gate and forgetting unwanted features. Subsequently, the output gate (O_t) regulates the exposure of the updated cell state as the output ($\mathbf{H}_t$). The hidden state ($\mathbf{H}_t$) is derived by applying the output gate to the hyperbolic tangent of the updated cell state, ensuring that the ConvLSTM captures and utilizes both short and long-term dependencies within sequential data, making it particularly effective for tasks involving spatial and volumetric information. The weight matrices ($W_{xi}, W_{hi}, W_{ci}, \ldots$) and bias vectors ($\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_c, \ldots$) associated with each gate are learned during the training process, enabling the ConvLSTM to adapt to the specific characteristics of the data it processes.

3.3 Regularized Feed Forward Neural Network

The Feed Forward Neural Network (FFNN) comprises two dense layers, with 512 and 256 neurons, employing the Gaussian Error Linear Unit (GELU) activation function, and includes a softmax layer for class labeling. To mitigate overfitting, each dense layer is accompanied by a dropout layer set at a rate of 0.5. Furthermore, weights in the fully connected layers are subjected to L_1 regularization, while biases undergo L_1L_2 regularization, providing additional protection against overfitting during the training process. These three components

are stacked and trained end-to-end through gradient centralization to make predictions, collectively referred to as the Spatial Sequence Attention Network. The combination of induced dropout, parameter regularization, and gradient centralization serves to prevent overfitting, allowing the model to generate meaningful predictions with reasonable accuracy. However, it is worth noting that the model's performance can vary depending on the pre-trained convolutional base used for morphological feature extraction.

3.4 Experiments & Results

3.4.1 Dataset

To validate the effectiveness of the proposed Spatial Sequence Attention network, we utilized a portion of the FBIRN dataset, which comprises both chronic SZ patients and control individuals (Keator et al. 2016). This dataset encompasses various data modalities, including structural MRI (sMRI), functional MRI (fMRI), single-nucleotide polymorphisms (SNP), and patient behavior information. However, we used just sMRI for our experiments to validate the effectiveness of the attention mechanism on morphological features. In total, this dataset consists of 186 participants, with 82 individuals diagnosed with SZ and 104 healthy subjects. Each MRI scan is downsampled by dimensions (121, 145, 121) for efficient processing.

3.4.2 Implementation Details

During experiments on this dataset, various hyperparameter values were explored, including learning rates (0.005 to 0.05), dropout rates (0.1 to 0.5), and regularization rates (0.001 to

0.1). Selected values to optimize model performance: initial learning rate (0.0001), batch size (32), training epochs (100), with L2 weight regularization (0.005) and L1_L2 bias regularization (0.005).

3.4.3 Quantitative Evaluation

For quantitative evaluation, we employ 5-fold cross-validation for SZ classification on the FBIRN dataset. We use standard classification metrics such as Accuracy, Precision, Recall, and F_1 Score to report the performance of conducted experiments. As a baseline model for our attention mechanism, we incorporate the existing Squeeze and Excitation Attention (SENet) into the feature extraction process of pre-trained DenseNet121 for the classification of SZ based on morphological MRI features. Subsequently, we aim to assess the effectiveness of the proposed Spatial Sequence Attention (SSANet) in terms of the feature extraction process of pre-trained DenseNet121. Additionally, we provide a comparative analysis with recent studies in the field that have utilized the same dataset and experimental configuration (Yu et al. 2022; Kanyal et al. 2023), presenting the results in Table 3.1.

Table 3.1 Comparison of Five-fold Cross-validation Metrics (in %) for SZ Classification

Method	Accuracy	Precision	Recall	F_1 Score
DenseNet (Huang et al. 2017)	70.00	-	-	-
XGBoost (Kanyal et al. 2023)	71.51	-	-	-
SENet	72.80 ± 0.05	77.84 ± 0.06	68.06 ± 0.08	66.79 ± 0.09
SSANet (Ours)	75.12 ± 0.06	77.22 ± 0.04	72.76 ± 0.08	71.91 ± 0.09

Looking at Table 3.1, it is evident that our proposed SSANet outperformed not only the Squeeze & Excitation Network (SENet) but also the fine-tuned DenseNet (Huang et al.

2017) and DenseNet features combined with XGBoost (Kanyal et al. 2023) in terms of accuracy. SENet, on the other hand, delivered competitive results. SSANet achieved higher accuracy and more balanced performance metrics. Additionally, the standard deviations accompanying our results highlight the stability of SSANet's performance. These findings underscore the efficacy of our proposed model in enhancing the SZ classification based on morphological sMRI features and emphasize its potential as a valuable addition to the field of medical image analysis and diagnostics.

3.4.4 Qualitative Evaluation

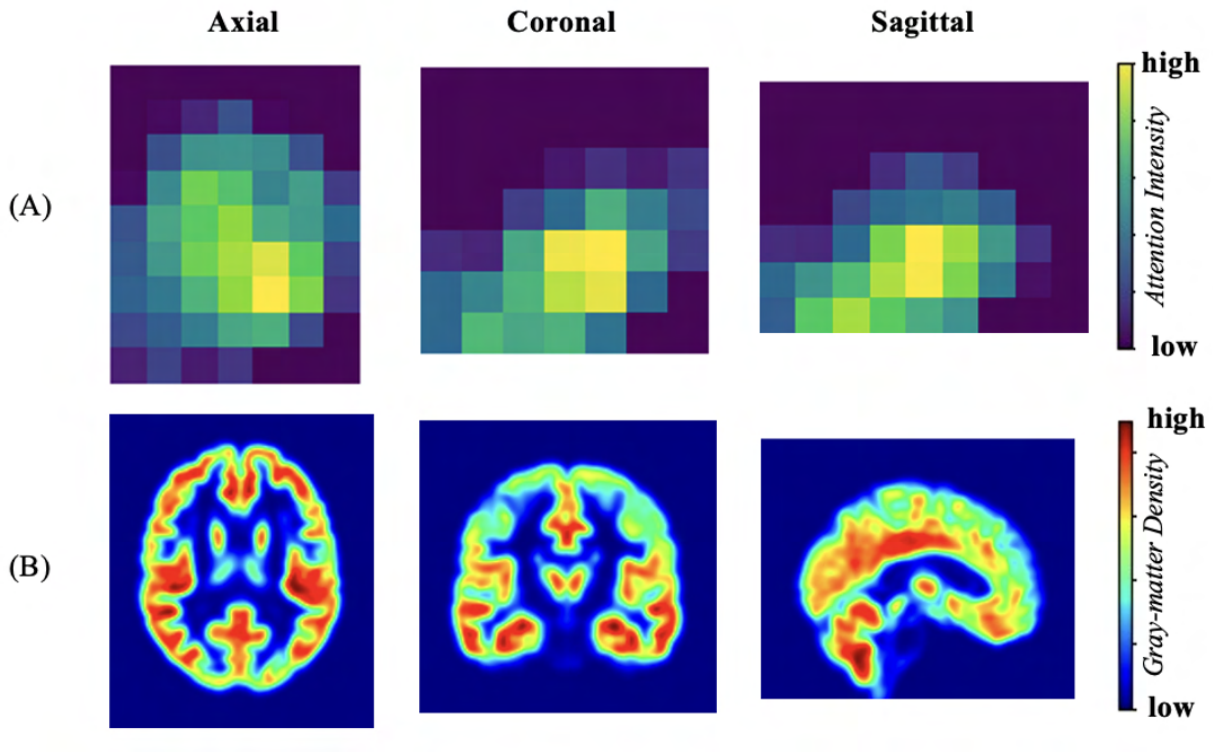


Figure 3.2 Visualization of (A) Attention Heatmaps; (B) Gray Matter Density Maps of a normal subject;

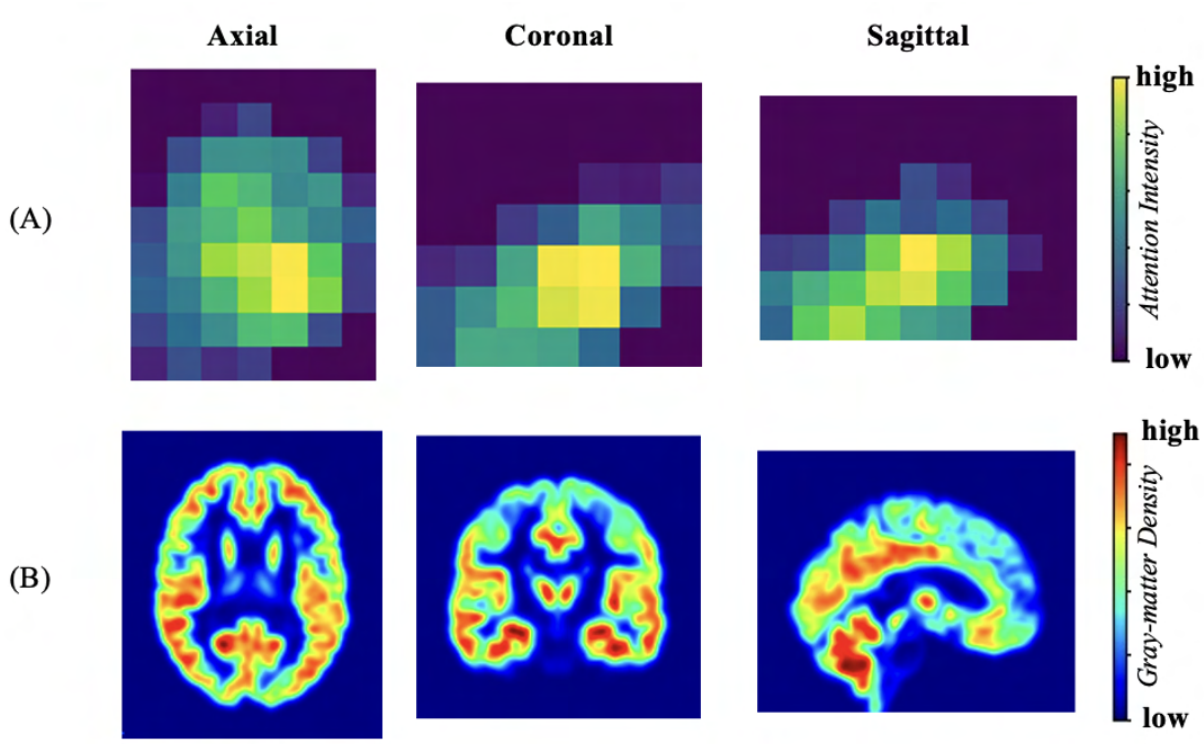


Figure 3.3 Visualization of (A) Attention Heatmaps; (B) Gray Matter Density Maps of a SZ subject; SSANet correctly gives attention around the disease impacted areas such as frontal lobe and temporal lobe. Note the decreased gray matter density in the SZ subject compared with the normal subject in those regions.

The qualitative evaluation involves visualizing attention heatmaps for axial, coronal, and sagittal slices. The superior performance of our proposed method for SZ classification is evident from the attention heatmaps from Figures 3.3 and 3.2, which can identify brain regions (frontal lobe, and temporal lobe) for attention and capture the gray matter density loss in SZ subjects. These attention maps contribute to enhanced discriminatory features, showcasing the effectiveness of our approach in capturing nuanced abnormalities associated with SZ.

CHAPTER 4
MULTI-MODAL IMAGING GENOMICS TRANSFORMER

This research presents our Multi-modal Imaging Genomics Transformer (MIGTrans), a deep learning model that combines structural brain imaging and functional connectome with genomic data using Cross-modal Multi-Head Attention (Vaswani et al. 2017) to distinguish between individuals with SZ and healthy controls (HCs). Figure 4.1 illustrates MIGTrans, while subsequent sections delve into its technical modules.

4.1 Genomic Connectome Representation Learning

4.1.1 Genomic Encoder

We define the learning of genomic representations through a neural network function $\phi(\cdot)$, which takes SNP, the genomic descriptor, $G \in \mathbb{R}^d$ as input and learns non-linear genomic representations $\bar{G} \in \mathbb{R}^{d'}$. This network consists of two dense layers with 2048 and 1536 units, activated by the Gaussian Error Linear Unit (GELU). Additionally, each layer is equipped with layer normalization and dropout mechanisms, ensuring robustness and preventing overfitting. The final layer outputs latent representations of the genomic features $\bar{G}$, encapsulating crucial information for discriminating SZ pathology.

$$G' = \text{LayerNorm}(\text{Dropout}(\Gamma(G \cdot W_1 + b_1), p_1)) \tag{4.1}$$

$$\bar{G} = \Gamma(\text{LayerNorm}(\text{Dropout}(\Gamma(G' \cdot W_2 + b_2), p_2)) \cdot W_3 + b_3) \tag{4.2}$$

Equations 4.1 and 4.2 illustrate the operations of $\phi(G)$ where G' denote intermediate

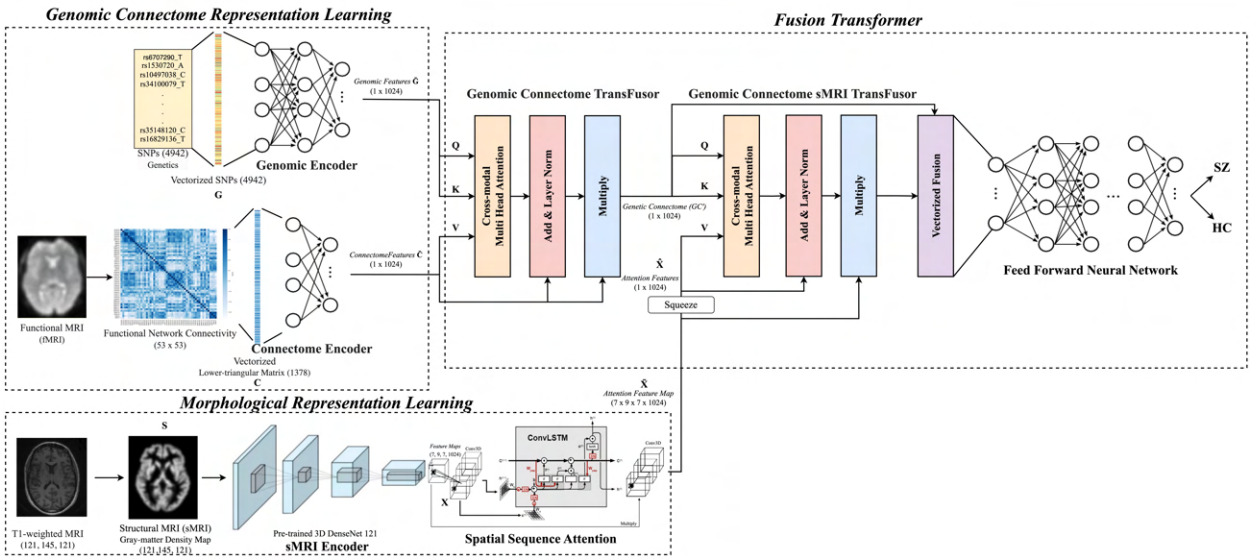


Figure 4.1 Architecture of Multi-modal Imaging Genomics Transformer (MIGTrans) - employing **Genomic & Connectome Encoders** for genomic & connectome representation learning; **Structural MRI encoder** and **Spatial Sequence Attention** for morphological representation learning; **Fusion Transformer** for multi-modal step-wise attentive integration

genomic representations processed by neural network layers. The GELU activation $\Gamma(x) = x \cdot \frac{1}{2}\left(1 + \frac{2}{\sqrt{\pi}} \int_0^{\frac{x}{\sqrt{2}}} e^{-\tau^2} d\tau\right)$ ensures non-linearity and preserves genomic complexities (Hendrycks & Gimpel 2016). W_1, W_2, and W_3 are weight matrices, and b_1, b_2, and b_3 are bias terms applied during linear transformations. p_1 and p_2 are control dropout probabilities.

4.1.2 Connectome Encoder

Similar to genomics, we define connectome representation learning through another neural network function $\psi(\cdot)$, which takes the functional connectome $C \in \mathbb{R}^f$ as input and learns non-linear connectome representations $\bar{C} \in \mathbb{R}^{d'}$. This function comprises a dense layer with 1536 units, activated by GELU, followed by layer normalization and dropout mechanisms. Subsequently, the output layer generates latent representations of the connectome, vital for

capturing intricate brain connectivity patterns.

$$C' = \text{LayerNorm}(\text{Dropout}(\Gamma(C \cdot W_4 + b_4), p_3)) \tag{4.3}$$

$$\bar{C} = \Gamma(C' \cdot W_5 + b_5)) \tag{4.4}$$

Equations 4.3 and 4.4 depict the processing steps of $\psi(C)$, where C' represents an intermediate connectome representation refined by the neural network layer, and $\bar{C}$ denotes the final output containing learned connectome features. $\Gamma(\cdot)$ denotes the GELU activation function. The weight matrices W_4 and W_5, along with bias terms b_4 and b_5, facilitate linear transformations essential for feature extraction. Additionally, the dropout probability p_3 regulates information flow, preventing overfitting, while layer normalization enhances the robustness of the network by standardizing input values.

4.2 Morphological Representation Learning

4.2.1 Structural MRI Encoder

We employ Transfer Learning paradigm and use a pre-trained 3D DenseNet121 to extract morphological features from gray matter density in sMRI images. This model transforms input sMRI S using a sequence of operations, including 3D convolutions, pooling, dense layers, and transition layers. The resulting output vector X encapsulates the subject's brain morphological features and is passed to a spatial sequence attention (SSA) module for learning attention across spatial and channel dimensions.

4.2.2 Spatial Sequence Attention

The SSA mechanism is tailored to capture spatial and channel dependencies within morphological feature maps X (Shaik & Cherukuri 2022a). Comprising a 3D convolutional (Conv3D) layer, a ConvLSTM layer, and another Conv3D layer, each component plays a crucial role in enhancing the feature maps (Chapter 3). The initial Conv3D layer extracts holistic features from the input, while the ConvLSTM, a variant of LSTM with convolutional operations, captures intricate spatial interactions and relationships across channels. By incorporating tanh and sigmoid activation functions, the ConvLSTM controls the flow of information and learns from sparse connections between input and state transitions, enabling it to learn both short and long-term dependencies within 3D feature maps (SHI et al. 2015). Finally, the output of Conv3D layer projects the refined features back to the original feature space. The evolution of the cell state (C_t) and hidden state (H_t) at each time step t is determined by input gate (I_t), forget gate (F_t), and output gate (O_t), considering both the current and past feature maps in the input sequence.

$$I_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} * C_{t-1} + b_i) \tag{4.5}$$

$$F_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} * C_{t-1} + b_f) \tag{4.6}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{4.7}$$

$$O_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} * C_{t-1} + b_o)H_t \tag{4.8}$$

The ConvLSTM operates based on fundamental equations 4.5 to 4.8, involving convolution ($*$) and element-wise product ($\odot$). The input gate (I_t) evaluates the importance of input feature maps (X_t) for inclusion in the cell state (C_t), while the forget gate (F_t) determines

the retention of features from the previous cell state (C_{t-1}). The cell state (C_t) is updated by selectively integrating input features and forgetting unwanted ones. Subsequently, the output gate (O_t) regulates the exposure of the updated cell state as the output (H_t). Through learned weight matrices ($W_{x_i}, W_{h_i}, W_{c_i}, \ldots$) and bias vectors ($b_i, b_f, b_c, \ldots$), the ConvLSTM adapts to the specific characteristics of the data, effectively capturing spatial and channel dependencies within morphological feature maps. Finally, SSA outputs attentive sMRI features $\bar{X}$.

4.3 Fusion Transformer

4.3.1 Genomic Connectome TransFusor

In this module, the learned genomic $\bar{G}$ and connectome $\bar{C}$ features are attentively integrated through Cross-modal Multi-Head Attention (x-MHA) with 2 heads, emphasizing the fusion of functional connectome from fMRI with non-imaging genomic biomarkers. Genomic features are transformed and passed as Query ($Q = Linear(\bar{G})$), Key ($K = Linear(\bar{G})$) and connectome features are passed as Value (V) to x-MHA. This strategic arrangement enables the model to compute attention between genomic and connectome features through a scaled dot-product mechanism, referred as Self-Attention (SA). Precisely, the attention scores are computed as the softmax of the scaled dot-product of the transformed genomic features, facilitating the selective focus on relevant interactions, and multiplying with connectome features. The resulting representation is combined with the original connectome features using layer normalization and element-wise addition. Further, the resultant features undergo

element-wise multiplication with the original connectome features, yielding the final fused connectome representation ($GC^{'}$).

$$SA(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d^k}})V \tag{4.9}$$

$$x\text{-MHA}(Q, K, V) = (\text{h}_1 \oplus ... \oplus \text{h}_i)W^o \quad \forall \text{h}_i = \text{SA}(QW_i^Q, KW_i^K, VW_i^V) \tag{4.10}$$

$$GC^{'} = \bar{C} \odot \text{LayerNorm}(\bar{C} + x\text{-MHA}(Q, K, V)) \tag{4.11}$$

Equations 4.9 to 4.11 illustrate the mathematical operations in scaled dot-product attention and x-MHA with two heads, as well as the formulation of the genomic connectome attention features. The symbols Q, K, and V represent matrices of queries, keys, and values, respectively. d^k denotes the dimensionality of the keys, while W_i^Q, W_i^K, W_i^V, and W^o are parameter matrices for linear transformation. The $\oplus$ signifies concatenation, $\odot$ indicates point-wise multiplication and the Softmax function is applied element-wise to normalize the attention scores in scaled dot-product attention. This approach is tailored to extract meaningful insights from the complex interplay between genomic and connectome data, enabling the model to effectively capture intricate dependencies specific to SZ.

4.3.2 Genomic Connectome sMRI TransFusor

Similar to the previous module, Genomic Connectome integration with sMRI is facilitated through x-MHA mechanism with 2 heads. This module aims to capture cross-modal relationships between imaging and non-imaging biomarkers. The genomic connectome features ($GC^{'}$) are passed as Query $\widetilde{Q}$ and Key $\widetilde{K}$, while the squeezed sMRI attention features ($\hat{X}$) are passed as Value $\widetilde{V}$ to the x-MHA. This strategic setup enables the model to selectively

attend to relevant regions in the sMRI features based on the genomic connectome attention features, thereby incorporating complementary information between modalities. The attention-weighted representations are fused with the original sMRI features through a process of layer normalization and element-wise addition. This fusion step ensures that the model can adaptively combine information from multi modalities while preserving the structural characteristics of the sMRI. Subsequently, the fused representations undergo further vectorized fusion through dense layers. This helps retain cross-modality features alongside fused representations, thereby improving the model's generalization performance.

$$GCS^{'} = \bar{X} \odot \text{LayerNorm}(\hat{X} + x\text{-MHA}(\widetilde{Q}, \widetilde{K}, \widetilde{V})) \tag{4.12}$$

Equation 4.12 elucidate the post-processing fusion after x-MHA, whose operations were illustrated in equations 4.9 and 4.10. The Feed Forward Neural Network acts as a classification head and utilizes the fused feature representations $GCS^{'}$ to predict the target labels $\hat{y} \in \{SZ, HC\}$. It consists of two dense layers with 512, 256 units each, leveraging GELU activation to extract intricate features and introduce non-linearity, enhancing the model's representational capacity. Employing layer normalization and a dropout mitigates overfitting, fostering improved model generalization by preventing excessive reliance on specific features. All the modules discussed above were trained end-to-end to minimize the cross-entropy loss, $\mathcal{L}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i))$, using the Adam optimizer for gradient optimization. The cross-entropy loss function $\mathcal{L}$ measures the dissimilarity between true labels y and predicted probabilities $\hat{y}$.

4.4 Experiments & Results

4.4.1 Dataset and Pre-processing

To prove the effectiveness of the proposed MIGTrans, we employed a subset of the Function Biomedical Informatics Research Network (FBIRN) dataset, containing sMRI, fMRI, and SNP data from 186 participants, including 82 SZ and 104 HC subjects (Keator et al. 2016). In our work, we select a subset of 4942 SNPs (one-hot encoded) related to SZ as identified by the psychiatric genomics consortium, extract the lower triangular matrix from (53×53) FNC matrix generated using Neuromark Atlas (Du et al. 2020), resulting in 1378 unique connections. Additionally, each sMRI is downsampled to $(121 \times 145 \times 121)$ dimensions & converted to gray matter density map using SPM12 (Kurth et al. 2015).

4.4.2 Implementation Details

We employ 5-fold cross validation to evaluate the effectiveness of proposed model. Throughout our experiments, various hyperparameter values were explored, including learning rates (0.005 to 0.05), dropout rates (0.1 to 0.5), and regularization rates (0.001 to 0.1). Selected values to optimize model performance: initial learning rate (0.001), dropout (p_1=0.5, p_2, p_3=0.3), batch size (16), training epochs (100), L2 weight regularization (0.005) and L1_L2 bias regularization (0.005).

4.4.3 Quantitative Evaluation

We conducted initial experiments to check the performance of single-modality approach using attention mechanism. We designed Self-Attention Networks for genomics (Genomic SANet) and connectome (Connectome SANet), and Spatial Sequence Attention Network (SSANet) for sMRI features. For multi-modal approaches, we explored various fusion strategies such as simple concatenation (Concat), Attentive Feature Fusion (AFF), SCCA, Att-DCCA for comparison with our proposed transformer-based method (Trans). As an ablation study, we also report the performance of the first-step fusion with genomics, connectome (GC) and the follow-up fusion with sMRI (GCS). Table 4.1 summarizes the classification performance of our proposed method and the baselines. The connectome alone performed best among single-modal features with 82.33%, supporting the SZ clinical finding (Kanyal et al. 2023). However, this is slightly lower than the 82.71% achieved by simply combining genomic and connectome attentive features.From this, it is clearly evident that multi-modal features are superior to when they are considered in isolation. In terms of multi-modal feature integra-

Table 4.1 Classification performance comparison of proposed model with baselines

Model	**Accuracy** (%)	**Precision** (%)	**Recall** (%)	F_1 **Score** (%)
Genomic SANet	63.64(±0.04)	64.48(±0.04)	60.77(±0.03)	59.28(±0.03)
Connectome SANet	82.33(±0.04)	81.87(±0.03)	82.04(±0.03)	81.88(±0.03)
sMRI SSANet	75.71(±0.06)	78.59(±0.04)	72.81(±0.08)	72.12(±0.09)
GC Concat	82.71(±0.05)	82.99(±0.04)	81.76(±0.04)	82.04(±0.04)
GC AFF	83.83(±0.04)	83.54(±0.03)	83.67(±0.04)	83.38(± 0.04)
GCTrans	84.82(±0.05)	84.45(±0.05)	84.56(±0.05)	84.45(±0.05)
GCS SCCA	84.29(±0.07)	83.86(±0.04)	83.99(±0.04)	83.92 (±0.05)
GCS Att-DCCA	84.31(±0.05)	84.56(±0.04)	84.02(±0.04)	83.88 (±0.05)
GCS Concat	84.32(±0.06)	84.55(±0.06)	83.46(±0.05)	83.79(±0.05)
GCS AFF	84.30(±0.05)	84.48(±0.06)	83.53(±0.04)	83.77(±0.05)
Proposed	**86.05(±0.02)**	**87.32(±0.03)**	**84.60(±0.02)**	**85.19(±0.02)**

tion, our proposed Trans method showed significant improvement in all metrics for two-way (GC) fusion when in comparison with concatenation and AFF, indicating 2.11%, 0.99% improvement. This trend continued even in three-way (GCS) fusion with 1.72% 1.73%, 1.74% 1.75% improvement when compared to SCCA, AFF, Att-DCCA and Concat in terms of accuracy. This highlights the benefit of our step-wise attentive integration of imaging with genomic biomarkers through cross-modal multi-head attention. Additionally, it is noteworthy that AFF showed a 0.02% lower accuracy compared to concatenation in three-way GCS fusion, indicating its struggle to maintain superiority as the number of modalities to integrate increases, given its reliance on linear fusion.

4.4.4 Qualitative Evaluation

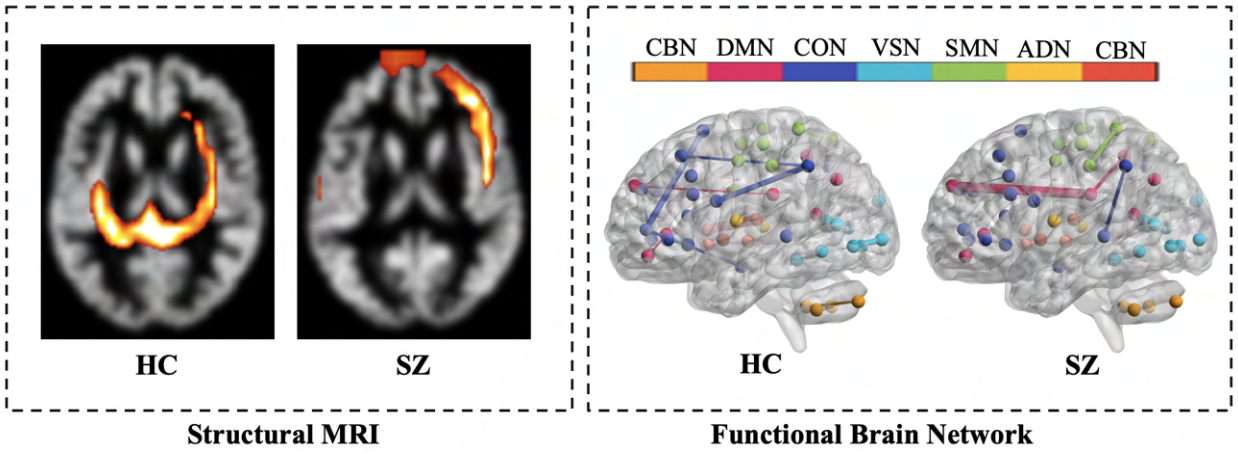


Figure 4.2 **Attention maps of sMRI** highlighting the frontal and temporal lobes in SZ samples, contrasting with thalamic regions in HC samples; **Functional brain networks** highlighting important connections in default mode (DMN) and cognitive control (CON) specific to SZ and HC compared with other subcortical (SCN), auditory (ADN), sensorimotor (SMN), visual (VSN), and cerebellum (CBN) networks.

To provide interpretation and clinical validation of our approach, we presented the atten-

tion maps of sMRI and important connections in functional connectome generated based on attention scores in Figure 4.2. For sMRI, attention scores from genome connectome sMRI TransFusor were multiplied with SSA Attention maps and visualized by GradCAM++. These maps identify primary regions highlighted in the frontal and temporal lobes in SZ samples, contrasting with thalamic regions in HC samples. This aligns with clinical observations indicating that individuals with SZ experience the most significant reductions in thalamic volumes and thickness within the frontal and temporal lobes (Haukvik et al. 2013). For functional brain network, we applied thresholding on FNC Attention scores to find top connectome features and observed denser connectivity patterns in HC subjects compared to individuals with SZ. In particular, the Default Mode Network (DMN) and Cognitive Control Network (CON) were revealed as significant interactions for SZ classification, aligning with previous study (Kim et al. 2010). Moreover, our approach successfully pinpointed the top three SZ-associated SNPs (*rs7643366, rs396911, rs13429970*), consistent with a genome-wide association study (European Bioinformatics Institute 2024; Kato et al. 2023). These results emphasize the discriminative potential of the attentive integration mechanisms in our MIGTrans.

CHAPTER 5

MULTI-MODAL MEDICAL TRANSFORMER

The goal of this research is to induce a novel deep learning framework, Multi-Modal Medical Transformer that attentively integrates clinical context from diagnostic keywords with visual insights obtained from multi-modal retinal scan images for comprehensive medical description generation. Figure 5.1 illustrates the complete architecture of proposed model. Technical specifications for each module are elaborated in subsequent subsections.

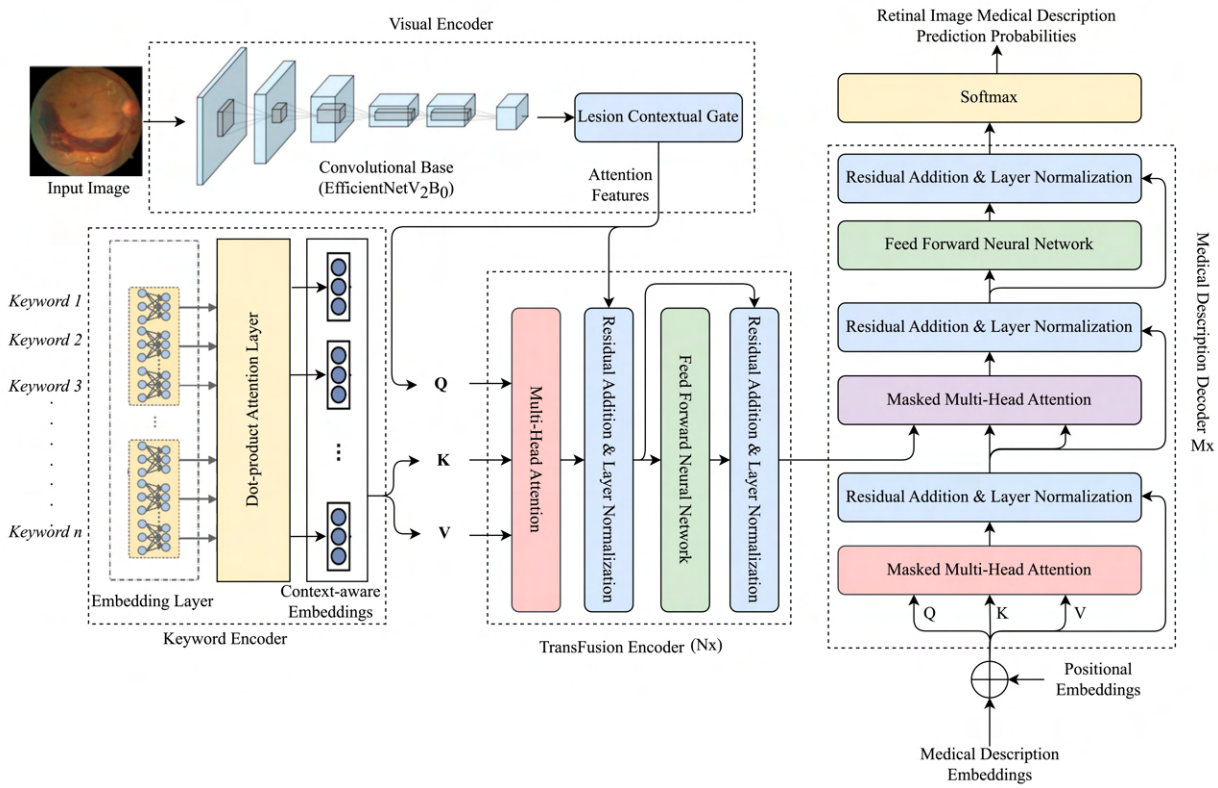


Figure 5.1 Architecture of proposed Multi-Modal Medical Transformer (M3T); **Visual Encoder** – learns attention-based representations from retinal images; **Keyword Encoder** – learns clinical-context embeddings from diagnostic keywords; **TransFusion Encoder** – integrates visual attention features and clinical-context embeddings, leveraging both visual and semantic information; **Medical Description Generation Decoder** – generates coherent and meaningful medical descriptions by attending to relevant visual and semantic cues, ensuring contextually appropriate and diagnostically relevant outputs;

5.1 Visual Encoder

The Visual Encoder is designed for converting the pre-processed retinal images into visual features through a special attention mechanism. This module is structured with a convolutional base and Lesion contextual gate operating collaboratively to extract lesion contextual visual representations.

5.1.1 Convolutional Base

The Convolutional base serves to extract initial visual representations capturing essential patterns in images. Convolutional Neural Networks (CNNs) excel in learning spatial representations across various image modalities, but training them from scratch requires substantial labeled data, often limited in medical image tasks like retinal scans. To overcome this limitation, we utilize pre-trained CNN models, initially trained on ImageNet, benefiting from generalizable visual features applicable to retinal images. Specifically, we choose EfficientNetV_2B_0 as the Convolutional base, which integrates Squeeze and Excitation (SE) blocks for attentive channel-wise information consideration, to extract spatial representations from multi-modal retinal images. Passing a retinal scan image X_R with dimensions $(356 \times 356 \times 3)$ through EfficientNet V_2B_0 yields visual features V of dimensions $(12 \times 12 \times 1280)$ from the final convolutional block.

5.1.2 Lesion Contextual Gate

The retinal representations obtained from the pre-trained EfficientNet V_2B_0 exhibit global information, where equal importance is assigned to each feature map. Despite the success

of attention mechanisms like Squeeze and Excitation blocks, our model still faces challenges in capturing lesion-specific contextual information crucial for accurate diagnosis. To address this limitation, we introduce a Lesion Contextual Gate Attention block that enhances retinal image representations with lesion context-rich information by using single attention block that considers global and local context information Cherukuri et al. (2024).

Initially, lesion context information is computed through global attention pooling which involves a point-wise convolution capturing essential features, and a softmax activation normalizes spatial elements generating an attention map. This attention map selectively focuses on relevant features in the initial spatial representations, providing lesion context information. Next, channel-wise dependencies through specialized point-wise convolution layers which process previously computed context information, introducing non-linearity and discerning intricate dependencies among different channels. The resulting features are aggregated with original spatial features, enriching contextual understanding within the retinal features. Finally, the gating function enhances the model's focus on lesion-specific regions where coefficients are computed based on global and local context information, selectively amplifying or suppressing features in retinal representations. The integration of weighted representations with normalization, and gating ensures effective highlighting of lesion-specific details, capturing both global and localized context features within the retinal images. Below are the mathematical computations involved in Lesion Contextual Gate module.

$$F_{gap} = \sum_{j=1}^{d} \frac{e^{W_c f_{r,j}}}{\sum_{m=1}^{d} e^{W_c f_{r,m}}} \qquad \forall f_{r,j} \in F_R, \tag{5.1}$$

where F_{gap} represents the global attention pooling of visual features F_R, W_c represents the the parameters of point-wise convolution. The numerator computes the exponential of the weighted visual features, and the denominator is the sum of exponentiated weighted visual features by using a softmax activation, resulting in a normalized global context feature.

$$F_c = F_R \oplus \sum_{i=1}^{d} W_2 \left(\mathrm{LN} \left(\Gamma \left(\sum_{j=1}^{k} W_1 f_{g_j} \right) \right) \right)_i \forall f_{g_j} \in F_{gap}, \tag{5.2}$$

where F_c represents the channel attentive feature incorporating global context, W_1, W_2 represents the parameters of point-wise convolutions, Γ is ReLU non-linearity and LN indicates Layer Normalization. The operation involves weighted summation and layer normalization of the global context features F_{gap}. This enriched features are then element-wise added to the original visual features F_R.

$$F_{att} = \sigma(W_\psi \cdot \Gamma(W_x F_R + W_g F_c + b_{xg}) + b_\psi), \tag{5.3}$$

where F_{att} represents the final attention-weighted visual features. This is computed by applying the gating mechanism using parameters $W_x, W_g, b_{xg}, W_\psi, b_\psi$. This gating is modulated by global context features from F_c and original visual features F_R.

5.2 Keywords Encoder

The Keywords Encoder functions as a language modeling encoder, transforming processed diagnostic keywords into context-aware embeddings. Within the encoder module, the Embedding layer takes the processed diagnostic keyword sequence as input, converting it into numerical values known as embeddings (Word2Vec). These embeddings, denoted as E, en-

capsulate individual keyword embeddings $ke_1, ke_2, ..., ke_n$, effectively capturing the semantic relationships embedded in the input keywords.

However, the initially generated embeddings are context-free, lacking explicit consideration of relationships between keywords. To address this limitation, the embeddings undergo processing by an Attention layer. This layer leverages a dot product operation to compute alignment scores between source (ke_s) and target (ke_t) embeddings, denoted as $\bar{ke}$ (Luong et al. 2015). These alignment scores reflect the importance or relevance of each keyword relative to others in the sequence. Subsequently, a Softmax activation function is applied to normalize the alignment scores, yielding attention weights ke_{att}. These weights furnish a context-aware representation of diagnostic keywords, assigning higher weights to more pertinent keywords and lower weights to less relevant ones. This mechanism enables the model to focus on the most informative keyword embeddings, facilitating the capture of crucial context and relationships between keywords. This context-aware representation is particularly valuable for the downstream task of Medical Description Generation.

$$\text{align_score}(ke_s, ke_t) = \bar{ke} = ke_s \odot ke_t, \tag{5.4}$$

$$ke_{att} = \text{Softmax}(W_{ke}, \bar{ke}) \qquad \forall ke_{att} \in KE_{att}. \tag{5.5}$$

Mathematically, Equation 5.4 expresses the computation of alignment scores through a dot product operation, while Equation 5.5 delineates the computation of attention weights using Softmax normalization, where W_{ke} denotes the parameters of the Softmax layer.

5.3 TransFusion Encoder

TransFusion Encoder is the backbone of proposed approach that is designed to receive two types of inputs, lesion contextual attention features and clinical context keyword embeddings from the Visual and Keyword encoders. Main objective of this module is to integrate the clinical context with visual insights from retinal images. It exploits the interactions between visual features and keywords by embedding the keywords using a self-attention mechanism, where attention weights are computed based on the interactions between the Query image (Q) and the Keywords Key (K), Value (V) pairs. This involves a dot product operation between the query and key embeddings, allowing the model to discern the importance of each keyword concerning the retinal image. The attention weights signify the degree of interaction between the image and each keyword, capturing their semantic relationships. Notably, higher attention weights emphasize keywords most relevant to the image, while lower weights assign less importance to less relevant keywords.

$$Q = W_q F_{att} \tag{5.6}$$

$$K = W_k K E_{att} \tag{5.7}$$

$$V = W_v K E_{att} \tag{5.8}$$

$$\text{Self Attention}(Q, K, V) = Z = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5.9}$$

$$Z_{norm} = \text{Layer Normalization}(Q + Z) \tag{5.10}$$

$$H = W_2 \Gamma(W_1, Z_{norm}) \tag{5.11}$$

$$F^{'} = \text{Layer Normalization}(H + Z_{norm}) \tag{5.12}$$

The mathematical operations underlying the computation of Multi-Head Attention weights Z are detailed in Equations 5.6 to 5.9. Here, W_q, W_k, W_v represent the parameters associated

with Attention Features and Context-aware Keyword Embeddings. The dot product operation in Equation 5.9 calculates keyword weights on the image query, with the result scaled by the length of the K vector (d_k). The softmax normalized ratio is then multiplied with V to produce attention probability weights. The subsequent operations involving Residual Addition, Layer Normalization (Equations 5.10 and 5.12), and Feed Forward Neural Network (FFNN) (Equation 5.11) contribute to refining the attention representations, resulting in $F^{'}$.

5.4 Medical Description Decoder

The Medical Description Generation Decoder is a language model based on Transformer Decoder architecture (Vaswani et al. 2017). It operates as a function of $(F^{'}, E_c, PE_c)$, representing Attention features, Clinical Description Embeddings, and Positional Embeddings (token embeddings + position embeddings), respectively. It comprises of two Masked Multi-Head Attention instances, the first enables self-attention, facilitating the model to focus on different segments of the input sequence while generating the clinical description. The second instance involves cross-attention, allowing the decoder to attend to the encoded retinal image features ($F^{'}$) and comprehend their relevance in generating subsequent words in the clinical description. During self-attention, a Masking (M) operation prevents the model from considering future words, ensuring generation based solely on preceding words, maintaining auto-regressive language generation. Following the attention mechanisms, a feed-forward neural network with two fully connected layers processes the attended and contextualized representations to generate the next word in the clinical description. To stabilize

the learning process, residual addition and Layer Normalization are applied post-attention and feed-forward neural network operations. Residual connections enable the model to bypass certain layers, retaining valuable information, while Layer Normalization normalizes outputs, enhancing training stability and overall decoder performance.

$$CE = E_c + PE_c \tag{5.13}$$

$$Z^{'} = \text{Multi Head Attention}(CE, CE, CE, \text{M}) \tag{5.14}$$

$$Z^{'}_{\text{norm}} = \text{Layer Normalization}(CE + Z^{'}) \tag{5.15}$$

$$Z^{''} = \text{Multi Head Attention}(F^{'}, Z^{'}\text{norm}, Z^{'}\text{norm}, \text{M}) \tag{5.16}$$

$$Z^{''}\text{norm} = \text{Layer Normalization}(Z^{''} + Z^{'}\text{norm}) \tag{5.17}$$

$$H^{'} = W_4\text{ReLU}(W_3, Z^{''}_{\text{norm}}) \tag{5.18}$$

$$R^{'}\text{final} = \text{Layer Normalization}(H^{'} + Z^{''}\text{norm}) \tag{5.19}$$

$$P = \text{Softmax}(R^{'}_{\text{final}}) \qquad P \in (p_0, p_1, p_2, ..., p_m) \tag{5.20}$$

Equations 5.13 to 5.20 encapsulate the mathematical expressions for generating clinical descriptions based on provided inputs. The resulting probability vector P signifies the probabilities of predicted words in the final clinical description.

The end-to-end training of the Multi-Modal Medical Transformer involves minimizing the cross-entropy loss function $\mathcal{L}(P, C)$, as represented in equation 5.21. This loss function quantifies the dissimilarity between predicted clinical description probabilities (P) and the actual clinical description (C). Summation over all vocabulary words (M), weighted by corresponding actual clinical description values (c_i), is performed. The logarithm of the predicted probability $\log(p_i)$ ensures more substantial errors are penalized accordingly.

$$\mathcal{L}(P, C) = -\sum_{i=1}^{M} c_i \cdot \log(p_i) \tag{5.21}$$

This loss function provides a training signal, guiding the model to minimize it, enhancing the capability to generate accurate medical descriptions for given retinal scan images.

5.5 Experiments & Results

5.5.1 Dataset

The DeepEyeNet dataset, introduced in (Huang et al. 2021c), is a large-scale collection of retinal images comprising 15,709 scans obtained from diverse imaging modalities such as Fluorescein Angiography (1,811 images), Fundus Photography (13,898 images), and Optical Coherence Tomography (OCT). Each image is meticulously annotated by expert ophthalmologists, providing multiple diagnostic keywords that capture medical observations and clinical descriptions related to the retinal findings. Keyword and clinical description lengths vary, with the longest keywords exceeding 15 words and the longest clinical descriptions surpassing 50 words. On average, both keywords and clinical descriptions have a word length ranging from 5 to 10 words. The dataset covers 265 distinct retinal diseases and symptoms, encompassing a broad spectrum of both common and less common conditions. This dataset is divided into standard splits: 60% for training (9,425), 20% for validation (3,142), and 20% for testing (3,142).

5.5.2 Pre-processing

Effective pre-processing is a foundation for optimizing the quality and interpretability of retinal images across diverse modalities such as color fundus, fundus auto-fluorescence, and OCT. In our approach, we implement crucial steps on acquired retinal scan images to mitigate variations in size, dimensions, and color arising from different devices and laboratory conditions. This includes resizing operations to standardize images to a uniform scale, as well as converting images into a three-channel format (Red, Green, Blue) to enhance processing flexibility and compatibility with pre-trained neural networks. Simultaneously, for clinical descriptions and diagnostic keywords, essential text processing steps are applied, encompassing removal of non-alphabetic characters, conversion to lowercase for uniformity, and handling rare words by replacing them with a UNK token. The list of provided keywords is transformed into a sequence by replacing comma separators with [SEP] during preprocessing.

5.5.3 Implementation Details

In our study, experiments were conducted on the DeepEyeNet dataset, adopting established settings from prior works (Huang et al. 2021c,b,a, 2022). To ensure comparability, consistent experimental conditions were maintained, and evaluation metrics (BLEU, CIDEr, ROUGE) aligned with those in referenced studies. The experiments were performed on an Nvidia P100 GPU (16GB Memory, 1.32GHz clock, 9.3 TFLOPS). Optimizing our model's performance involved carefully chosen hyperparameters, with a learning rate of 0.004, Adam optimizer, batch size of 64, and dropout rate of 0.2. The model architecture comprises a

two-layer fully connected feed-forward neural network (FFCN) with hidden layer sizes of 512 and 256, demonstrating tailored efficiency. A vocabulary size of 5000, 300-dimensional word embeddings, and set sequence lengths (5 to 50 words) contribute to the model's effectiveness. Categorical cross-entropy as the loss function solidifies our hyperparameter choices for successful model training.

5.5.4 Quantitative Evaluation

Table 5.1 Results demonstrating the effect of incorporating Keywords as input to the model and integrating Lesion Contextual Gate with Visual features, along with Context-aware Attention to Keyword Embeddings;

Image	Image Attn	Kwds	Kwds Attn	BLEU Scores				Other Metrics	
				BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	ROUGE
✓	✗	✗	✗	0.114	0.067	0.054	0.026	0.296	0.197
✓	✓	✗	✗	0.184	0.110	0.068	0.032	0.316	0.230
✓	✓	✓	✗	0.214	0.152	0.105	0.072	0.349	0.296
✓	✓	✓	✓	0.394	0.312	0.291	0.208	0.537	0.493

In the ablation study in Table 5.1, we meticulously dissect the components of our M3 Transformer to evaluate their individual and cumulative impacts on medical description generation for retinal images. Notably, the progressive inclusion of visual attention, keywords, and attention mechanisms on both visual and keyword features leads to an improvement in BLEU scores. The results signify the benefit of visual attention in focusing on relevant regions within retinal images, while the incorporation of diagnostic keywords significantly enhances the model's ability to integrate clinical context. The culmination of these components in the full model demonstrates a substantial leap in BLEU scores, reinforcing the synergistic effect of jointly considering visual and clinical information for robust medical

description generation.

Table 5.2 Comparative Study of Recent Best Models with Proposed Context Gating Vision Transformer trained on DeepEyeNet Dataset, highlighting the impact of incorporating keyword embeddings and Attention on Visual features;

Model	BLEU Scores				Other Metrics	
	BLEU@1	BLEU@2	BLEU@3	BLEU@4	CIDEr	ROUGE
DeepOpth	0.184	0.114	0.068	0.032	0.361	0.232
Deep Context Encoding Network	0.219	0.134	0.074	0.035	0.398	0.252
Contextualized Keywords	0.203	0.142	0.100	0.073	0.389	0.211
Non-local Attention	0.230	0.150	0.094	0.053	0.370	0.291
M3 Transformer	**0.394**	**0.312**	**0.291**	**0.208**	**0.537**	**0.493**

In the comparative study illustrated in Table 5.2, our proposed M3 Transformer Network is pitted against recent leading models in the field of retinal image clinical description generation. Against models like DeepOpth, Deep Context Encoding Network, Contextualized Keywords, and Non-local Attention, our M3 Transformer exhibits superior performance across all BLEU scores, showcasing its capacity to generate more accurate and coherent clinical descriptions. These results highlight the distinctive strength of our approach in seamlessly integrating multi-modal information from retinal images and clinical keywords, making significant strides in addressing the challenges of generating clinically relevant descriptions for diverse retinal conditions.

5.5.5 Qualitative Evaluation

In Figure 5.2, we provide a visual representation of the Lesion Contextual Gate Attention heatmaps and corresponding overlays, showcasing the explainability of our proposed approach. The heatmaps highlight the visual insights extracted from diverse retinal imaging modalities, including color fundus, gray fundus, optical coherence tomography (OCT), and

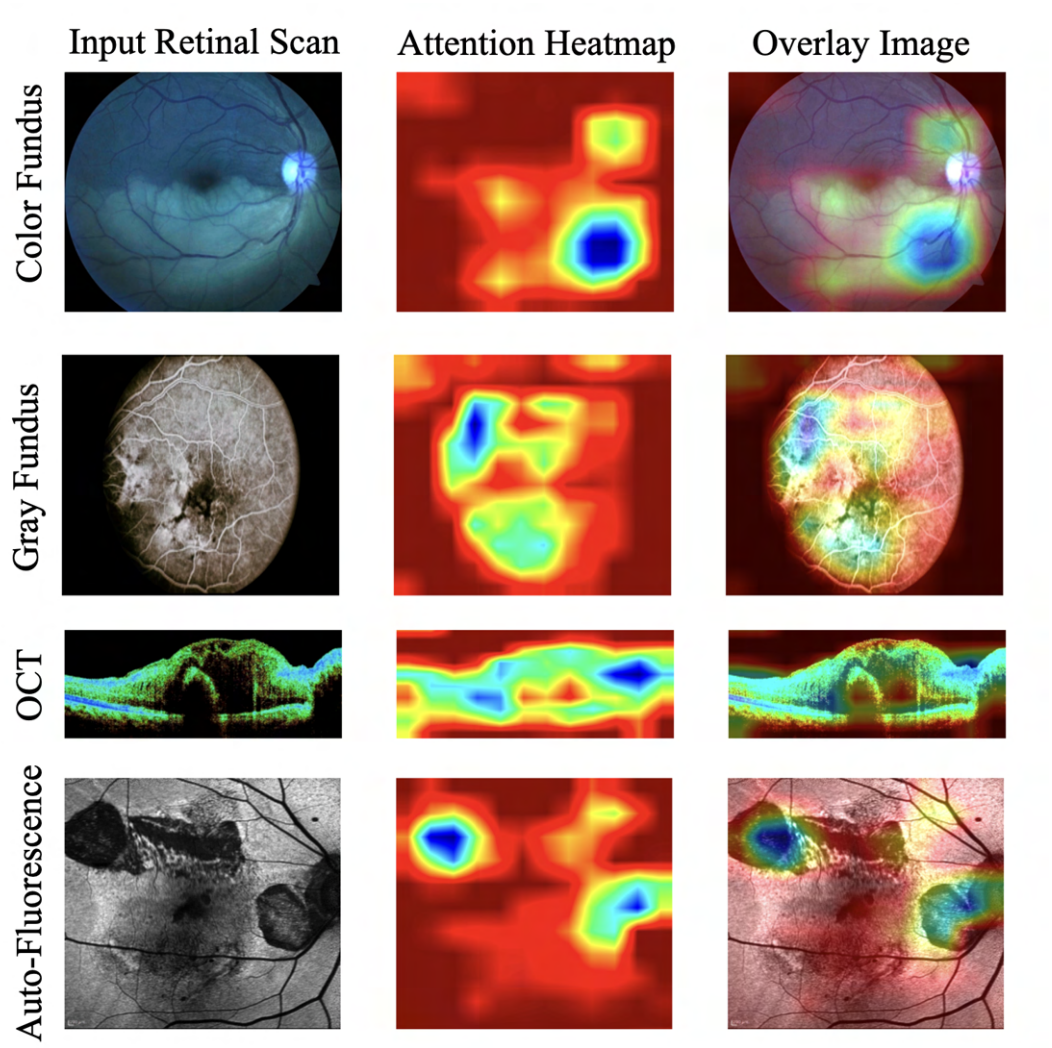


Figure 5.2 Visualization of Lesion Contextual Gate Attention heatmaps and corresponding overlays, highlighting the visual insights from color fundus, gray fundus, OCT, and Auto-Fluorescent retinal scan images

autofluorescent retinal scan images. The Lesion Contextual Gate effectively focuses on relevant regions within the retinal images, emphasizing areas crucial for clinical understanding. This visualization underscores the model's capability to integrate information from various modalities, providing ophthalmologists with interpretable indications of the visual cues influencing the generated clinical descriptions.

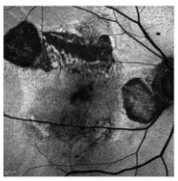
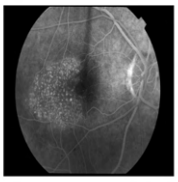
Retinal Image	Keywords	Groundtruth	Non-Local Attention [16]	M3 Transformer
	autofluorescence imaging[sep]age-related macular degeneration (amd)	an autofluorescence image of a 78 year old man with an age related macular degeneration on his both eyes	Autofluorescence to image of the right eye of a [age] year old [gender] with acute decrease in vision mainly right eye is with pigment clumping and optic nerve drusen in the right eye	autofluorescence image of a 76 year old woman with age related macular degeneration on her right eye and the left eye
	idiopathic macular telangiecta-sia[sep]parafoveal telangiecta-sia[sep]juxtafoveal telangiectasis	the fellow eye was unremarkable on this red free image	The telangiectasis occurs unilaterally in the macula be of choroidal folds peripheral nevus shows cnvm in this eyes shows resolved with myopia over the macula and lung remained and remained over the exam was remained cnvm. Resolved and pdt.	the telangiectasis occurs unilaterally in the temporal half of the macula in an area of 1–2 disc diameters the anomalies are noted on this red free image in this early frame of the angiogram

Figure 5.3 Comparison of Actual, Non-Local Attention, and M3T medical descriptions

Figure 5.3 offers a qualitative comparison of actual, Non-Local Attention (Huang et al. 2022), and predicted medical descriptions using retinal images and keywords as input. Our M3 Transformer outperforms the Non-Local Attention model by generating more accurate and coherent clinical descriptions. For instance, in the first case, the Non-Local Attention model produces a description with generalities about age and gender, while our M3 Transformer provides specific details about macular hole conditions in both eyes. Similarly, in the second case, our model accurately identifies and describes the presence of idiopathic macular telangiectasia, parafoveal telangiectasis, and juxtafoveal telangiectasis, offering detailed insights into the retinal anomalies. This comparison highlights the superior performance of our proposed approach in capturing nuanced clinical information and generating clinically relevant descriptions with greater accuracy and specificity.

CHAPTER 6
CONCLUSION

This thesis has demonstrated the transformative potential of attentive multi-modal learning in medical diagnostics, focusing on schizophrenia (SZ) classification and automated medical report generation for retinal images. By integrating diverse modalities—such as structural and functional MRI, genomic biomarkers, and clinical text—this research has advanced precision medicine, addressing key challenges in diagnostic accuracy, interpretability, and efficiency. The Spatial Sequence Attention Network (SSANet) improved sensitivity to subtle brain changes in SZ, while the Multi-modal Imaging Genomics Transformer (MIGTrans) provided interpretable insights into structural, functional, and genetic markers. The Multi-Modal Medical Transformer (M3T) enabled precise and coherent medical report generation, overcoming challenges of data scarcity and complex pathology interpretation. These frameworks prioritize interpretability and scalability, offering scalable solutions for healthcare.

6.1 Future Research Directions

This research opens several promising avenues for future exploration to enhance multi-modal learning in medical diagnostics. Key directions include:

- **Scalable Parallel Modality Fusion:** Developing frameworks for simultaneous integration of multiple modalities, improving computational efficiency and flexibility for large-scale, heterogeneous data.
- **Multi-modal Latent Distribution Learning:** Exploring shared representations

across modalities using techniques like VAEs or GANs to enhance fusion efficacy and interpretability.

- **Robust Handling of Missing Modalities:** Designing adaptive frameworks to address missing data, ensuring reliable performance in real-world clinical scenarios.
- **Zero-shot and Few-shot Learning:** Advancing techniques for low-resource settings, enabling diagnostics with minimal labeled data through Zero/Few-shot learning.

These extensions will bridge the gap between visual data and clinical context, driving innovation in personalized, data-driven healthcare solutions.

REFERENCES

Beddiar, D.-R., Oussalah, M., & Seppänen, T. 2022, Artificial Intelligence Review, 1

Bi, Y., Abrol, A., Fu, Z., & Calhoun, V. 2023, in 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), IEEE, 1–5

Bodapati, J. D., Shaik, N. S., & Naralasetti, V. 2021, Signal, Image and Video Processing, 15, 923

Cannon, T. D., Rosso, I. M., Hollister, J. M., Bearden, C. E., Sanchez, L. E., & Hadley, T. 2000, Schizophrenia bulletin, 26, 351

Chan, Y. H., Wang, C., Soh, W. K., & Rajapakse, J. C. 2022, Frontiers in Neuroscience, 16, 866666

Chen, X. et al. 2023, Biomedical Signal Processing and Control, 80, 104293

Cherukuri, T. K., Shaik, N. S., Bodapati, J. D., & Ye, D. H. 2025, in 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 1–5

Cherukuri, T. K., Shaik, N. S., & Ye, D. H. 2024, in 2024 IEEE International Conference on Image Processing (ICIP), IEEE, 3098–3104

Chuang, C.-Y., Lin, Y.-T., Liu, C.-C., Lee, L.-E., Chang, H.-Y., Liu, A.-S., Hung, S.-H., & Fu, L.-C. 2023, IEEE Transactions on Neural Systems and Rehabilitation Engineering

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., & Barnard, K. 2021, in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 3560–3569

Du, L., Liu, K., Yao, X., Risacher, S. L., Han, J., Guo, L., Saykin, A. J., & Shen, L. 2018,

in 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 356–361

Du, Y. et al. 2020, NeuroImage: Clinical, 28, 102375

European Bioinformatics Institute. 2024, GWAS Catalog, https://ebi.ac.uk/gwas/, accessed: February 7, 2024

Goldblatt, R. et al. 2024a, medRxiv, 2023

——. 2024b, Nature Mental Health, 1

Haukvik, U. K., Hartberg, C. B., & Agartz, I. 2013, Tidsskrift for Den norske legeforening

Hendrycks, D., & Gimpel, K. 2016, arXiv preprint arXiv:1606.08415

Herdade, S., Kappeler, A., Boakye, K., & Soares, J. 2019, Advances in neural information processing systems, 32

Hu, J., Shen, L., & Sun, G. 2018, in Proceedings of the IEEE conference on computer vision and pattern recognition, 7132–7141

Hu, M., Qian, X., Liu, S., Koh, A. J., Sim, K., Jiang, X., Guan, C., & Zhou, J. H. 2022, Schizophrenia research, 243, 330

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017, in Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708

Huang, J.-H., Wu, T.-W., & Worring, M. 2021a, in Proceedings of the 2021 International Conference on Multimedia Retrieval, 645–652

Huang, J.-H., Wu, T.-W., Yang, C.-H. H., Shi, Z., Lin, I., Tegner, J., Worring, M., et al. 2022, in Proceedings of the IEEE/CVF winter conference on applications of computer

vision, 1606–1615

Huang, J.-H., Wu, T.-W., Yang, C.-H. H., & Worring, M. 2021b, in 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 3762–3766

Huang, J.-H. et al. 2021c, in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2442–2452

Hung, C.-C., Lin, K.-H., & Chang, H.-A. 2024, Medicina, 60, 2060

Jafri, M. J., Pearlson, G. D., Stevens, M., & Calhoun, V. D. 2008, Neuroimage, 39, 1666

Kamal, A. H., Jishan, M. A., & Mansoor, N. 2020, in 2020 International Conference on Decision Aid Sciences and Application (DASA), IEEE, 822–826

Kanyal, A., Kandula, S., Calhoun, V., & Ye, D. H. 2023, in 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), IEEE, 1–5

Kato, H., Kimura, H., Kushima, I., Takahashi, N., Aleksic, B., & Ozaki, N. 2023, Journal of Human Genetics, 68, 175

Keator, D. B. et al. 2016, Neuroimage, 124, 1074

Kim, D. I. et al. 2010, Neuroinformatics, 8, 213

Kim, S., & Misra, A. 2007, Annu. Rev. Biomed. Eng., 9, 289

Kurth, F., Gaser, C., & Luders, E. 2015, Nature protocols, 10, 293

LeCun, Y., Bengio, Y., et al. 1995, The handbook of brain theory and neural networks, 3361, 1995

Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., & Chang, X. 2023, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3334–3343

Luong, M.-T., Pham, H., & Manning, C. D. 2015, arXiv preprint arXiv:1508.04025

Misra, I., & Maaten, L. v. d. 2020, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6707–6717

Park, H. W., Kim, S. Y., & Lee, W. H. 2023, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 626–636

Radford, A. et al. 2021, in International conference on machine learning, PMLR, 8748–8763

Sapkota, R., Thapaliya, B., Ray, B., Suresh, P., & Liu, J. 2025, Information, 16, 160

Sapkota, R., Thapaliya, B., Suresh, P., Ray, B., Calhoun, V. D., & Liu, J. 2024, in ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2071–2075

Schnack, H. G., Nieuwenhuis, M., van Haren, N. E., Abramovic, L., Scheewe, T. W., Brouwer, R. M., Pol, H. E. H., & Kahn, R. S. 2014, Neuroimage, 84, 299

Shaik, N. S., & Cherukuri, T. K. 2021, Machine Vision and Applications, 32, 126

——. 2022a, Applied Intelligence, 52, 15105

——. 2022b, Computers in Biology and Medicine, 141, 105127

Shaik, N. S., Cherukuri, T. K., Calhoun, V., & Ye, D. H. 2024a, in 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 1–5

Shaik, N. S., Cherukuri, T. K., Calhoun, V. D., & Ye, D. H. 2025a, in 2025 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 1–5

Shaik, N. S., Cherukuri, T. K., Veeranjaneulu, N., & Bodapati, J. D. 2024b, Machine Vision and Applications, 35, 73

Shaik, N. S., Cherukuri, T. K., & Ye, D. H. 2024c, in 2024 IEEE International Conference on Image Processing (ICIP), IEEE, 3037–3043

Shaik, N. S., Veeranjaneulu, N., & Bodapati, J. D. 2025b, Machine Vision and Applications, 36, 56

SHI, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., & WOO, W.-c. 2015, in Advances in Neural Information Processing Systems, Vol. 28 (Curran Associates, Inc.), 802–810

Thapaliya, B., Akbas, E., Chen, J., Sapkota, R., Ray, B., Suresh, P., Calhoun, V. D., & Liu, J. 2025a, Medical Image Analysis, 101, 103433

Thapaliya, B. et al. 2025b, Medical Image Analysis, 103462

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. 2017, Advances in neural information processing systems, 30

Wismüller, A., Vosoughi, A., Kasturi, A., & Hadjiyski, N. 2023, in Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging, Vol. 12468, SPIE, 278–286

Wismüller, A., & Vosoughi, M. A. 2021, in Medical Imaging 2021: Computer-Aided Diagnosis, Vol. 11597, SPIE, 337–350

World Health Organization. 2023, Blindness and Visual Impairment, accessed: 2025-01-20

Wu, T.-W., Huang, J.-H., Lin, J., & Worring, M. 2023, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1859–1868

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. 2015, in International conference on machine learning, PMLR, 2048–2057

Yu, H., Florian, T., Calhoun, V., & Ye, D. 2022, in 2022 IEEE International Conference on Image Processing (ICIP), IEEE, 3291–3295

Zheng, Y., He, M., & Congdon, N. 2012, Indian journal of ophthalmology, 60, 428

Zhou, R., Zhou, H., Chen, B. Y., Shen, L., Zhang, Y., & He, L. 2023, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 681–691

A Papers Published

This appendix lists the research papers published as part of this thesis. The following publications are included:

1. **Paper 1**: *Shaik, Nagur Shareef*, Cherukuri, Teja Krishna, Calhoun, Vince, and Ye, Dong Hye. "Spatial Sequence Attention Network for Schizophrenia Classification from Structural Brain MR Images." In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. Shaik et al. (2024a).

2. **Paper 2**: *Shaik, Nagur Shareef*, Cherukuri, Teja Krishna, Calhoun, Vince D, and Ye, Dong Hye. "Multi-modal Imaging Genomics Transformer: Attentive Integration of Imaging with Genomic Biomarkers for Schizophrenia Classification." In *2025 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025. Shaik et al. (2025a).

3. **Paper 3**: *Shaik, Nagur Shareef*, Cherukuri, Teja Krishna, and Ye, Dong Hye. "M3T: Multi-Modal Medical Transformer to Bridge Clinical Context with Visual Insights for Retinal Image Medical Description Generation." In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3037–3043. IEEE, 2024. Shaik et al. (2024c).

These publications form the core contributions of this thesis and have been peer-reviewed and accepted in reputed conferences.