

A Hybrid Deep Learning Approach with ConvNeXt and Vision Transformer for Detecting Schizophrenia in fMRI Scans

by

Afif Rayhan Pranto
21301425

Mahin Abdullah
21301359

Raheek Muhammad Raiyan
21301282

Saadman Muhib
24341111

Saimum Reza Siam
21101164

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025

© 2025. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



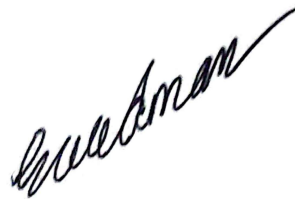
Afif Rayhan Pranto
21301425



Mahin Abdullah
21301359



Raheek Muhammad Raiyan
21301282



Saadman Muhib
24341111



Saimum Reza Siam
21101164

Approval

The thesis/project titled “A Hybrid Deep Learning Approach with ConvNeXt and Vision Transformer for Detecting Schizophrenia in fMRI Scans” submitted by

1. Afif Rayhan Pranto (21301425)
2. Mahin Abdullah (21301359)
3. Raheek Muhammad Raiyan (21301282)
4. Saadman Muhib (24341111)
5. Saimum Reza Siam (21101164)

Of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)



Mr. Md. Sabbir Ahmed
Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Jawaril Munshad Abedin
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Schizophrenia being one of the most complex mental disorders disrupts the default livelihood of the affected along with their close ones. Thus posing significant challenges to the individuals and the healthcare system worldwide as they often fail to diagnose the illness on time for effective treatment and management. In this case, MRI scans offer a promising opportunity for diagnosis but interpreting these complex images requires advanced analytical techniques which all physicians might not have. This study proposes a novel hybrid deep learning model composed of ConvNeXt and Vision Transformer (ViT) architectures to classify schizophrenia from 3D functional MRI. ConvNeXt is used for local feature extraction by capturing detailed and hierarchical features from the MRI images. Meanwhile, Vision Transformers make the model better at seeing the bigger picture and understanding the complex connections in the data by using a special self-attention process. Utilizing the subtle neurobiological markers in fMRI, the hybrid model will facilitate fast and proper detection of schizophrenia patients. This model demonstrates immense potential for automated schizophrenia detection by combining advanced deep learning techniques to develop precise and reliable diagnostic tools for mental healthcare.

Keywords: Schizophrenia, Magnetic Resonance Imaging, fMRI, Vision Transformer model, ViT, ConvNeXt

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, we are grateful to our thesis Supervisor Mr. Md. Sabbir Ahmed sir for his kind support and advice in our work. He helped us whenever we needed help.

Thirdly, to our Co-supervisor Jawaril Munshad Abedin for her kind support and advice in our work. Her continuous support, guidance, and insightful feedback have greatly contributed to the depth and quality of our research.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
1 Introduction	1
1.1 Introduction to Schizophrenia	1
1.2 Introduction to Magnetic Resonance Imaging (MRI)	1
1.3 Introduction to Machine Learning and Computer Vision	3
1.4 Research Problem	4
1.5 Research Objectives	4
2 Literature Review	6
3 Methodology	16
3.1 ConvNeXt	16
3.2 Vision Transformer	17
3.3 Work Plan	18
3.4 COBRE Dataset	19
3.4.1 Dataset Collection	19
3.4.2 Dataset Preprocessing	19
3.5 OpenNeuro UCLA Dataset	20
3.5.1 Dataset Collection	20
3.5.2 Dataset Preprocessing	20
3.5.3 Dataset Augmentation	21
3.6 ConvNeXt Architecture	21
3.7 Vision Transformer Architecture	25
3.8 Hybrid Architecture	30
3.8.1 ConvNeXt Output and Feature Projection	30
3.8.2 Transformer Encoding and Output Extraction	30
3.8.3 Class Token Extraction	31
3.8.4 Final MLP Classifier and Output Generation	31

4	Result and Analysis	33
4.1	Training and Evaluation Protocol	33
4.2	Evaluation Metrics	36
4.3	Hybrid Model Analysis	37
4.4	Comparing Model Analysis	40
4.5	Performance Summary Table	42
4.6	Confusion Matrix Analysis	43
4.7	Best Epochs per Fold	46
4.8	Comparative Discussion	47
5	Limitations and Future Work	52
6	Conclusion	53
	Bibliography	57

Chapter 1

Introduction

1.1 Introduction to Schizophrenia

Schizophrenia (SZ), a severe and long-lasting mental illness, is characterized by abnormalities in thought, reality perception, emotional reactions, and social conduct. Although each person experiences the condition differently, its symptoms can significantly disrupt day-to-day functioning, posing serious personal, social, and financial difficulties for patients, their families, and larger healthcare systems. Schizophrenia is ranked eighth among the leading cause of disability-adjusted life years around the world which affect an estimated 1% of the population [32]. It is one of the most widespread mental disorders that significantly affects an individual's ability to work and learn [9]. Approximately 24 million people, or 1 in 300 individuals (0.32%) worldwide, are diagnosed with schizophrenia, with the rate increasing to 1 in 222 people (0.45%) among adults. Traditional beliefs about the origins of schizophrenia vary widely, with common explanations including demon possession, curses, bewitchment, evil spirits, the evil eye, divine will, sorcery, and punitive justice [1]. Compared to the general population, people with schizophrenia are two to three times more likely to die early. This is due to their increased susceptibility to a range of other illnesses, including physical, metabolic, infectious, and cardiovascular diseases [6]. When compared to other chronic mental and physical health conditions, the financial burden of schizophrenia is exceptionally high. This includes not only the direct costs of healthcare but also indirect costs associated with social services, lost productivity, involvement with the criminal justice system, and other non-healthcare factors [4].

1.2 Introduction to Magnetic Resonance Imaging (MRI)

In the field of medical diagnosis, a lot of imaging techniques and devices have been developed with the advancement of science and technology. One such device is MRI, which stands for Magnetic Resonance Imaging. It is an effective non-invasive technique for mapping the body's interior structure and some of its functional characteristics. Using nonionizing electromagnetic radiation, is believed to have no exposure-related risks. By utilizing radio frequency (RF) radiation and well regulated magnetic fields, it creates excellent 3D pictures of the body in any of plane.

The subject is placed inside a huge magnet, which creates an external relatively safe strong magnetic field to create the MR image. As a result, many atoms in the body, especially hydrogen, align their nuclei with the magnetic field, which is then used to apply an RF signal. The body releases energy, which a computer detects and uses to create the MR image [2]. Different kinds of images are produced depending on the time and intensity of the signals that are released, and each includes unique features that make it useful for identifying different disorders. In the case of a brain MRI, T1-weighted imaging (T1) focuses on the anatomy of the brain by producing high contrast between tissues such as white matter (WM) and gray matter (GM). It frequently serves as a baseline reference for other MRI sequences and offers exceptional resolution of structural characteristics. On the other hand, T2-weighted imaging (T2) is especially helpful in identifying edema or swelling, which may be a precursor of diseases like tumors or inflammation, given that it displays fluid-filled areas like cerebrospinal fluid (CSF). Contrast-enhanced T1-weighted imaging, or T1C, uses a contrast agent, such as gadolinium, to make blood vessels, inflammation, and lesions more visible. These can reveal blood-brain barrier abnormalities, which are frequently seen in schizophrenia and other mental diseases. FLAIR (Fluid-attenuated inversion recovery) is a specialized sequence that inhibits signals from the cerebrospinal fluid, improving the detection of lesions close to the ventricles. It also helps in the detection of anomalies near the fluid-filled spaces of the brain. The ability of structural MRI (sMRI) scans to identify macroscopic anomalies makes them an indispensable component of clinical imaging services. Because of their ease of implementation, they can be used in non-specialized areas of research. One possible drawback is that structural alterations might take place at a comparatively late point in the pathological progression that results in psychosis. Furthermore, structural MRI has limited resolution, just like all imaging methods, and certain structures cannot be divided into their individual parts. Because of these drawbacks, detecting a structural alteration does not always reveal important details about the underlying process of psychosis development, and caution should be exercised when interpreting them. [14].

To solve the problem presented by the limitations of structural MRI, functional MRI or fMRI should be considered. It is a powerful neuroimaging technique similar to structural MRI but more powerful. This is because fMRI works by detecting changes in blood oxygenation and flow that occur in response to neural activation. On the other hand, structural MRI provides only the static images of the brain at a certain time. Functional magnetic resonance imaging (fMRI) allows the investigation of functional processes by capturing dynamic changes in neural activity with respect to time. One of the most popular fMRI methods known as BOLD (Blood Oxygen Level-Dependent) imaging uses the fluctuations in the magnetic properties of hemoglobin when it becomes oxygenated as a measurement. This particular technique allows physicians and experts to identify specific regions of the brain that are involved in various cognitive and sensory tasks. The most common application of fMRI is resting-state fMRI (RS-fMRI) where brain activity is recorded while the subject is not engaged in any specific task but typically lying quietly with eyes closed or fixated. Among the various types of fMRI, resting state fMRI is the most common as the patients do not need to perform any task. Moreover, this type of fMRI is more likely to be found among all the datasets while having the ability to reveal

deep-level alterations that may not be clear in task-based fMRI. However, fMRI data are sensitive to motion artifacts and physiological noise. Furthermore, its relatively low spatial and temporal resolution may limit the detection of subtle abnormalities. At the same time, the pros of fMRI outweigh the cons in using it as a tool for the diagnosis of both mental and neurological abnormalities. [5]

1.3 Introduction to Machine Learning and Computer Vision

With the recent explosion in popularity of Artificial Intelligence (AI), the world is becoming more used to it day by day as we continue to use it for our daily menial tasks. Likewise researchers are integrating machine learning (ML) and computer vision (CV) in their respective fields. They have revolutionized the medical field by discovering diagnosis techniques of complex illnesses that were not possible until now. Machine learning can be used to learn patterns using the data obtained, while computer vision can analyze pictorial data precisely and effectively to make decisions. These techniques can be applied to neuroimaging data to uncover subtle and blurry features that may have been missed by traditional clinical analysis. This is especially valuable in disorders like schizophrenia that we discussed earlier.

Various machine learning and computer vision techniques have been developed in the last decade as a powerful tool for medical imaging. Convolutional neural networks (CNN) [3] have been widely used due to their ability to capture detail with very limited data efficiently. While vision transformers (ViT) have shown great results when provided with sufficient data [18]. The combination of these architectures opens up the possibility of extracting not only local anatomical details but also spatial context from MRI volumes. This is crucial in identifying the complex alterations within schizophrenia affected patients. These models can learn unique features that classify individuals with schizophrenia from healthy controls by analyzing brain volume, connectivity patterns or signal fluctuations.

In conclusion, we can say that the possible benefits of applying machine learning methods to schizophrenia diagnosis are enormous. Besides, this will benefit the medical sector by helping the physicians with automated systems for faster and easier diagnosis. Moreover, these systems can be trained on large datasets in the future to continuously improve over time. This particular thesis introduces a novel approach in schizophrenia diagnosis despite challenges such as limited data availability to help those who are suffering due to this life altering illness.

1.4 Research Problem

Magnetic Resonance Imaging (MRI) is one of the most advanced techniques used in the diagnosis of neurological disorders. The complicated mental illness known as schizophrenia can be recognized by abnormalities in thought processes, emotional responsiveness, and social behavior, and it manifests through subtle alterations in brain structure. Abnormalities in both structure and function can be identified because of the precise images of the brain that an MRI gives. However, because brain abnormalities are complex and schizophrenia manifests differently in each person, it is still difficult to precisely diagnose the condition using magnetic resonance imaging (MRI). Even with improvements in medical imaging technologies, the capacity of existing techniques to discern significant patterns from these intricate MRI data is frequently compromised. Various machine learning models have been researched to improve this situation. In each research, we have been seeing significant improvements in accurately identifying disorders like schizophrenia. We are trying to improve the accuracy rate even more using the combination of two models namely ConvNeXt and Vision Transformers. This hybrid model will be tested on MRI datasets and we hope to improve the accuracy and to reduce the computational burden which is common in deep learning models. This research aims to fill the gap in the current research by providing a more robust and efficient approach to schizophrenia detection using better deep-learning approaches.

1.5 Research Objectives

- **To develop a hybrid deep learning model combining ConvNeXt and Vision Transformers for schizophrenia detection using brain MRIs:** The goal is to combine the methods of ConvNeXt and Vision Transformers to form a new hybrid model. We will do so by collecting and evaluating other studies conducted in this particular field. This will help us to identify the advantages and disadvantages of the previous models. Thus we can adjust our model accordingly to our particular needs and create a more accurate and reliable method to detect schizophrenia from MRIs.
- **To evaluate the accuracy and robustness of the hybrid model compared to existing models:** Another of our goal is to compare our hybrid model with the existing traditional models to figure out the accuracy and functionality between them. Comprehensive testing will help us to evaluate whether the novel hybrid model will offer better accuracy in detecting schizophrenia than other models or not.
- **To assess the computational efficiency of the proposed hybrid model:** Analyzing the model's computational requirements, such as processing speed and memory consumption, is critical. If the model becomes too complex or heavily dependent on computation power, it will result in delays. As a result valuable time of both doctors and patients will be lost. This objective ensures that the hybrid model is effective and practical for deployment in real-world clinical situations where computational resources may be limited.

- **To illustrate the usefulness of the hybrid model in real-world diagnostic scenarios involving schizophrenia detection:** For this objective, we will use real-world and diverse datasets to verify the model's accuracy in the detection of schizophrenia. We will try out non-experimental and non-controlled datasets to get proper results by validation.
- **To contribute to the advancement of medical image analysis techniques through the integration of cutting-edge architectures:** Our paper will contribute to the rapidly growing field of using computer vision in the medical sector. Future researchers will further the advancements we made based on our research. As a result, better hybrid models will be developed to detect not just schizophrenia, but also other neurological diseases.

Chapter 2

Literature Review

Zhang et al. (2023) used structural MRI data to address the issue of identifying severe mental illness (SMI) before it manifests, including schizophrenia, bipolar disorder, anxiety disorder, and depressive disorder [34]. Before this, there were no real-world models for early and accurate SMI detection. Multiple Instance Learning (MIL) is a poorly supervised convolutional neural network (CNN) technique. This algorithm was suggested by the authors to train a model that successfully captured subtle structural brain anomalies associated with SMI. To standardize the data, they preprocessed the MRI scans before implementing the MIL model. 2D image data from MRI were used as input instead of 3D data to save computational power before applying the MIL algorithm to differentiate between SMI patients and healthy individuals. The model achieved an accuracy of 0.76 with a sensitivity of 0.77, specificity of 0.74, and AUC of 0.82, indicating strong classification performance. Yet when evaluated on other independent datasets using different imaging techniques, the model's performance fell. This indicates limitations with regard to data generalization. Additionally, the specificity of the model was lower than its sensitivity, suggesting a higher rate of false positives. Moreover, their research indicated that the accuracy of MRI-based diagnostic models for psychiatric disorders is comparatively lower. Finally, the authors suggested more validation on bigger and more varied datasets for the model to be more effective.

In this paper [15], the authors addressed the difficulties of using structural MRI data to diagnose schizophrenia. Traditional approaches in the medical industry frequently rely on neurologists' subjective clinical assessments, which leads to uneven diagnosis results. They suggested a deep learning technique to find patterns in brain MRI scans linked to schizophrenia that is based on a 3D convolutional neural network (CNN). The authors used information from five publicly available datasets—BrainGluSchi, COBRE, MCICShare, NMorphCH, and NUSDAST—to carry out this. MRI scan images underwent preprocessing and were transformed into a video format. A video input was created by converting each slice from the 3D NIFTI format to a frame and then combining these frames. The video format was used so that the model could mimic the way clinicians read brain images in day-to-day life. The deep learning model was trained accordingly to avoid overfitting along with other necessary adjustments. The study reported an AUC of 0.96 and an overall accuracy of 97% for distinguishing schizophrenia patients from healthy controls, which is higher than that of previous studies. With an AUC of 0.72, the model's

performance significantly dropped when used on a new clinical dataset (Uijeongbu St. Mary’s), suggesting that it had trouble generalizing across various demographics and imaging techniques. Furthermore, other mental illnesses that share similar characteristics with schizophrenia were not mentioned in the study for comparison.

Chen et al. (2020) focused on the challenge of identifying abnormal brain regions in schizophrenia using structural MRI [12]. To reduce complexities, they chose only paranoid schizophrenia and normal controls from the COBRE (Centers for Biomedical Research Excellence) dataset. The study addressed the problem of the “curse of dimensionality,” where the high number of features from MRI scans could lead to overfitting in machine learning models. Thus, images were segmented into three tissue probability maps (TPMs), including GM, WM, and cerebrospinal fluid (CSF). The authors proposed a hierarchical feature selection approach, combining two-sample t-tests and recursive feature elimination (RFE) with a support vector machine (SVM) classifier. This reduced the feature space while retaining the most relevant brain region information for classification. Implemented on the COBRE dataset, the method achieved an accuracy of approximately 85% and an AUC of 0.97 ± 0.03 for white matter features. While for grey matter it showed an accuracy of approximately 79% and an AUC of 0.90 ± 0.06 . Given these favorable findings, the study was limited by a small sample size and lacked comparison with deep learning models, which are becoming more and more popular in the interpretation of medical images. Furthermore, this study did not take into account related mental conditions like bipolar disorder, which have many similarities. To further increase classification accuracy, the scientists recognized the necessity of including functional MRI data and using deep learning methods.

In a very recent study [39], the problem of using structural MRI to classify brain tumor grades was looked at by Mehmood and Bajwa. This problem is similar to the detection of other severe illnesses, such as schizophrenia, in that it requires accurate identification of subtle anomalies in the brain. Using a transfer learning method alongside a next-generation convolutional neural network, the ConvNext architecture, was the proposed solution. Utilizing the previously preprocessed BraTS 2019 dataset, which contained a variety of MRI sequences (such as T1, T2, T1C, and FLAIR), a pre-trained ConvNext model was fine-tuned for execution. After feeding the CNN three distinct MRI sequences (T1, T1C, and FLAIR), together at the same time, the model was able to classify low-grade and high-grade gliomas with a 99.5% classification accuracy. While for an individual sequence, T1C gave the best accuracy. Despite its high accuracy, the study had limitations regarding the small dataset size and the heavy reliance on transfer learning, which may not always generalize well to new clinical data without further fine-tuning.

Montalbo et al. (2024), presented a complete analysis of different kinds of Deep Learning models in order to evaluate their productivity while they are being used to diagnose brain tumors from MRI reports [40]. It presents a comparative analysis of Vision Transformer models against pure CNN models. Here, the researchers proposed an easy and usable hybrid model to yield better results from MRI scans. To solve the problem, lightweight versions of the models were used by analyzing them through different metrics like- AUC scores, accuracy, etc. Moreover, consistent hyperparameters were used due to budget constraints. With an accuracy of 95.65%,

ViT performed the best out of all. On the other hand, MobileViT proved to be the more cost-friendly option. On the contrary, Swin Transformer had the lowest performance across criteria, but ConvNeXt performed well in situations of meningioma. However, for transformers, a huge amount of computational resources is needed. For this reason, hybrid models have been suggested to be used for detecting such tumors.

Hamran et al. (2023), addressed the challenge of detecting brain tumors from MRI images due to the significant delays in the diagnostic report generation and also the complexity of brain tumor identification [30]. Therefore, the use of CNNs has been proposed for brain tumor classification since it has a high-efficiency rate in capturing spatial hierarchies in image data. That's why skip connections are introduced to improve accuracy, which ultimately addresses the issue of accuracy degradation that keeps occurring in deep networks. In order to use CNN models, the given MRI datasets are pre-processed to use them as training data. This is done by resizing the images and converting them into grayscale. Different transfer learning techniques are used where ResNet or EfficientNet are fine-tuned for brain tumor classification. According to the results, CNN-based models get an accuracy level of over 98% on test datasets. Both transfer learning techniques and skip connections have contributed to the improvement of performance in deeper networks, improving classification accuracy and also ensuring proper information flow across all the layers. However, there are a lot of limitations of CNNs in brain tumor detection. Firstly, these models are computationally expensive and might prove to be too costly for medical implications in a limited environment. Also, MRI datasets of brain tumors are not available publicly which is quite a challenge.

In this study [22], the authors showed a comparative analysis between hybrid and pure transformer models for brain tumor segmentation by using multi-modal MRI data. CNN is mostly suited to capture features in a small region, whereas transformers are more helpful for detecting long-range features overall. So, it is suggested here that integrating ViT with CNN backbones will result in better accuracy of the results. In this research, PyTorch and MONAI were implemented and 5-fold cross-validation was used in order to ensure a proper evaluation. Moreover, along with hybrid CNN-ViT models, and pure Transformer models, in total eight models were trained. Despite higher accuracy and better evaluation, hybrid models tend to prove more costly than other transformer models.

Liu et al. (2024), in their research [38], analyzed the failings of CNNs in capturing long-range features and for transformers, there is a struggle with the local details. Therefore, this resulted in an inaccurate medical image segmentation during disease detection. Here, in order to solve these problems a dual-branch hybrid model which integrates both CNNs and Transformers with attention mechanisms that will focus on the major parts of the images to generate better accuracy. It is called the Global-Local Fusion Network (GLFUnet). Here, Swin Transformer encoders and ConvNext are used to generate the hierarchical features. Furthermore, using the AtFF(Attention Feature Fusion) module the features of ConvNeXt and Swin Transformer encoders were integrated. The research that was conducted on the provided datasets, resulted in the GLFUnet having much better performance in all aspects. However, GLFUnet requires extensive training data, due to which there are chances

of facing scalability challenges and also generalization in diversified medical images.

Khan (2024) [37] addressed the challenges and complexities that the traditional methods faced while diagnosing Alzheimer’s disease (AD) due to the complex structure of the brain using MRI scans. In order to tackle this problem, a unique feature map advancement technique is used that combines a Residual Convolutional Neural Network (CNN) with a Transformer model. Residual CNN is initially used to capture the local features and then transformer mechanisms were used to extract features for long-range dependencies. The proposed model achieved great diagnostic performance compared to the norm, which resulted in improved accuracy and sensitivity. The only limitations it mostly faced are- generalizability of different findings and also faced challenges in clinical settings.

Odusami et al. (2023) [31] provided a detailed study on the early detection of a neurodegenerative disorder affecting cognitive functions and memory, which is Alzheimer’s disease (AD). They presented a multimodal fusion approach through ViT or Vision Transformers which was optimized through transfer learning. The study tackles the crucial problem of identifying AD at the late mild cognitive impairment (LMCI) and early mild cognitive impairment (EMCI) stages by merging sMRI and PET imaging data. The method used to align MRI and PET data using Procrustes analysis and preprocess them using anisotropic diffusion filtering is described in detail in the publication along with for feature extraction the usage of VGG16 and Discrete Wavelet Transform (DWT) for image decomposition. A ViT model, which has been refined on the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset, is used to classify AD. Using an accuracy of 93.75%, the model’s performance using PET data is especially remarkable, greatly outperforming that of single-modality models. When compared to using MRI or PET data alone, multimodal fusion dramatically improves diagnosis accuracy, according to one of the paper’s main findings. The promise of ViTs in medical imaging applications, especially in the context of multimodal fusion, is demonstrated by the Vision Transformer’s superior performance over conventional CNN-based models. Smaller datasets, lack of optimization of fusion parameters as well as the absence of more imaging modalities like attention-based networks might have limited the accuracy, generalizability, and performance. To sum up, this study shows how ViTs and multimodal fusion can be used to diagnose Alzheimer’s disease early, which makes it a viable area for further study.

Alshayegi [35] in this article proposed a meta-analysis of machine intelligence techniques and perceptive systematic review for Alzheimer’s disease (AD) identification. It thoroughly assesses both conventional machine learning models (e.g., CNN, SVM, RNN, CNN + LSTM) and Vision Transformers (ViTs) highlighting the latter’s superior ability to identify Alzheimer’s disease phases. By using various algorithms, the study is supposed to tackle early AD identification including the progression of Dementia stages by using MRI images. The usage of Kaggle’s Alzheimer’s dataset, consisting of 6400 T1-weighted MRI brain pictures categorized into four different stages of dementia, is a major strength of this study by determining the condition of the cortical gray matter and ventricles of the brain. How well the medium-sized ViT ‘viT_base_patch16_224’ model pre-trained on ImageNet-21k captures both global and local brain properties is demonstrated by the thorough use of Vision Transform-

ers. ViTs significantly outperform conventional CNNs in the classification of various stages of Alzheimer’s disease due to their capacity to analyze images as patches and make use of self-attention mechanisms. With an accuracy of 99.83% along with specificity and sensitivity of 99.88% and 99.69% respectively which shows the high reliability of identifying AD by capturing the abnormalities of brain tissues and dependencies of brain regions effectively. The comparatively limited dataset of four AD stages and the requirement of substantial computational resources for ViTs are the limitations that might be improved for expanding advanced dementia phases. Exploring modalities like PET scans and the improvement of model interpretability can also be options. To sum up, this paper offers strong proof of Vision Transformers’ effectiveness in identifying Alzheimer’s disease, but it also urges more research into multimodal integration and model improvement for better diagnostic abilities.

Bi et al. (2023) presented an innovative method for predicting schizophrenia by combining functional MRI (fMRI) and structural MRI (sMRI) data into a Multimodal Vision Transformer (MultiViT) model [29]. The model combines functional and structural brain imaging, improving upon conventional unimodal approaches and producing more accurate and comprehensible results by utilizing a cross-attention mechanism. The main issue discussed is the difficulty of categorizing schizophrenia because of the intricate interactions between functional and structural changes in the brain. Better prediction accuracy and a more thorough understanding of the brain are made possible by this creative synthesis of modalities (fMRI and sMRI data). The study examined two datasets, from different hospitals of China and US hospitals, including healthy controls and patients with schizophrenia. In order to manage structural and functional connectivity (FNC) matrices, the data was processed using Vision Transformers (ViTs). A cross-attention mechanism merged the two data sets for more interpretability and prediction power. Moreover, the usage of 3D ViTs for sMRI and 2D ViTs for FNC matrices and then fusing the two modalities using cross-attention mechanism for the information on functional and structural domains generating the attention maps to highlight the parts of the brain that are most likely to predict schizophrenia have outperformed the CNN based models as it is only limited to local features unlike the processing of MRI images as patches by MultiViT for capturing long-range dependencies. Thus, the research gave an outstanding accuracy of 83.1% along with F1 and AUC of 0.840 and 0.833 respectively which outperformed the multimodal and unimodal baselines and key brain areas linked to schizophrenia, including the hippocampus, precentral gyrus, and anterior cingulum were revealed by the model’s attention maps. Despite this accuracy, the limitations might be the usage of less amount of datasets as MultiViT needs large amounts of data sets along with the usage of imaging modalities like PET scans would improve the detection of disease. In summary, the MultiViT model provides a strong and comprehensible approach to the prediction of schizophrenia, showing notable advancements over conventional models and offering a more profound understanding of the link between the structural and functional aspects of the brain. Future studies will concentrate on adding more varied data sources to the model in order to increase its scalability and efficiency.

In this study [36], in this paper, explore the complex relationship between the structure and function of the brain in schizophrenia context using a unique conditional

generative adversarial network (cEViT-GAN) model. The authors focused on the link between FNC or functional network connectivity that is obtained from functional MRI (fMRI) data and the changes in grey matter volume recorded by sMRI. It has been done by integrating various data sources using ViTs along with self-attention mechanisms to address the problem of finding biomarkers for schizophrenia. The key factor is the generation of FNC matrices from GM data utilizing the blockwise self-attention mechanism of 3D ViTs capturing local and global brain interactions. With the data sets from the hospitals of the US and China including schizophrenia patients and healthy controls. Key brain areas like dorsolateral prefrontal cortex (DL-PFC) and medial prefrontal cortex (mPFC) were identified in the study which are essential for comprehending the structural-functional abnormalities in schizophrenia by using 3D ViTs with blockwise self-attention mechanism for processing MRI images, generating the FNC matrices from 3D GM patches and using pre-trained ViT patch embeddings, the model’s efficiency is increased and then transformed into FNC maps using a multi-layer perceptron (MLP). The results are outstanding with a 97% correlation for FNC matrices and a Pearson correlation of 0.74 between produced and real FNC data. Compared to traditional CNN-based models, the cEViT-GAN outperforms them by employing self-attention to enhance feature extraction and interpretation of complex brain structures. The paper also emphasizes the model’s blockwise attention mechanism, which reduces computation complexity without reducing effectiveness. The study does, however, point out some drawbacks, including the computational resources needed and the dataset’s concentration on schizophrenia. Future research could incorporate more data sources, such as EEG or PET scans, and broaden the model to encompass other neurological conditions. To sum up, the cEViT-GAN model provides a strong framework for identifying biomarkers for schizophrenia and offers comprehensible information on the areas of the brain impacted by the illness. Its effectiveness and generalizability to different mental health issues may be further enhanced by future studies.

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that causes the patient to have different social and behavioral patterns than other people. The complexity of ASD makes it hard to get an accurate diagnosis, particularly in the medical sector where time is important. To solve this problem, Khadem-Reza and Zare [24] in their paper used machine learning for automatic diagnosis using structural MRI data (sMRI). One researcher showed the use of cortical surface thickness as a classifier for ASD, while another researcher utilized a multiparameter classification technique with SVM, achieving high sensitivity. More recently, deep learning models, such as the multilevel deep network used by a researcher, have further improved accuracy in detecting ASD. In the present study, the authors used several machine learning classifiers: SVM, RF, KNN, and ANN, using features like cortical thickness, sulcus depth, and white and gray matter volume. Their findings showed that the artificial neural network (ANN) achieved the highest accuracy of 88.46%, outperforming other models like SVM and KNN. However, the small sample size and the presence of comorbidities limited the study’s generalizability. For future improvements, larger datasets should be used and the use of fMRI could enhance diagnostic accuracy in future research. Overall, machine learning models show promising improvements for the automatic diagnosis of ASD, though more detailed data and stronger algorithms such as hybrid models are needed to make

the results better and faster.

In their paper, by utilizing cutting-edge image processing techniques and machine learning approaches, Marghalani et al. [10] have improved the diagnostic accuracy of brain pathology classification, especially for illnesses like brain tumors and Alzheimer’s disease (AD). The main challenge has been the accurate detection and classification of these conditions from MRI images due to their complex and varied appearances. The researchers have proposed solutions that use feature extraction methods and classifiers. Bag of Features (BoF) and Support Vector Machine (SVM) classifiers can be mixed to identify brain tumors, AD, and normal brain images, and this method focuses on high accuracy. The methodology has four key stages: MRI image acquisition, pre-processing for noise reduction, feature extraction using BoF, and classification with SVM. The results have shown a high classification accuracy of around 97% across different trials, demonstrating the method’s effectiveness in handling multi-class classification challenges. The limitations include the computational complexity of feature extraction techniques like SURF and the requirement for additional grid size tuning for better performance. Even though these challenges exist, the proposed solution marks a significant step forward in automating brain pathology detection, with the potential for further enhancements in computational efficiency and feature selection techniques.

Parkinson’s Disease(PD) is a mental disease that can make people decline mentally and affect their daily lives. This can be temporarily solved by medicines but an accurate diagnosis is needed for better treatment. With traditional diagnostic methods, there is a result of 25% misdiagnosis. The recent use of image processing and machine learning has been providing promising results in PD detection. Models like CNN, DenseNet, and ResNet have shown significantly better performance in PD classification using MRI images. These suffer in easy use in clinical environments. In the paper of Palakayala and Kuppusamy [41], the use of Attention-based models has been brought up. This model introduces methods that focus on critical regions in medical images, thus enhancing feature extraction. Hybrid models that combine CNNs with attention mechanisms have further improved accuracy in detecting brain abnormalities. The proposed model, AttLUNet, achieved an accuracy of 99.58%, outperforming LeNet-5 and U-Net. The f1-score, recall, and precision were all 99.59%, demonstrating superior performance in the detection of Parkinson’s Disease. Due to the use of a single dataset, the study is quite limited. In conclusion, while deep learning models such as AttentionLUNet offer significant improvements in PD detection, future work must address the need for multimodal data integration and enhanced model understanding so that these can be used easily in a clinical environment.

The growing complexity of Alzheimer’s disease (AD) diagnosis has led to the need for advanced classification models to effectively identify and differentiate between its various stages. In this paper by Ajagbe et al. (2021) [16], the primary challenge addressed is the multi-classification of Alzheimer’s disease using medical images. The researchers suggested addressing the multi-classification issue and enhancing diagnostic accuracy by utilizing three deep-learning models: CNN, VGG-16, and VGG-19. The models were developed through the use of Python programming packages, specifically TensorFlow and Keras, on a publicly available dataset. The

models were assessed using metrics including accuracy, AUC, F1-score, precision, recall, and computational time. The findings showed that VGG-19 outperformed CNN in terms of F1-score (50.04%), recall (50.04%), and computational time (0.86 hours), while VGG-19 performed best in terms of accuracy (77.66%), AUC (81.55%), and precision (58.48%). The limitations included the lengthy computation times of the VGG models and the public datasets. The public datasets might not be accurate or they might not reflect the brain-related issues that people face in real world. To solve this issue, the researchers suggested the use of diverse datasets and powerful resources such as GPU. To sum up, while the proposed deep learning models show potential for AD classification, more improvements in processing efficiency and dataset quality are needed for a larger variety of clinical uses.

According to this study by Wang et al. (2021) [20], it can be difficult to diagnose brain disorders like schizophrenia and ADHD using MRI pictures. Neuroimage classification performance suffers from inadequate accounting for brain illnesses across many scales by the available approaches. The authors proposed a 3D Multiscale View Convolutional Neural Network with Attention (3D MVA-CNN). This model is used to figure out multi-scale brain connectivity disorders using convolutional kernels of varying sizes. Then it enhances feature importance dynamically using an attention mechanism. The solution was implemented using a combination of ResNetXt and Squeeze-and-Excitation (SE) mechanisms. The Multiscale View (MV) module applies parallel convolutional layers with varying kernel sizes to capture different brain connectivity scales. An attention mechanism dynamically assigns importance to the features extracted from these layers. Compared to VGGNet, ResNet, and Inception-V3, the 3D MVA-CNN model performs way better on the datasets used for schizophrenia and ADHD. It gives better performance compared to the previous techniques in the identification of brain disorders, with 78.8% accuracy for ADHD and 88.2% accuracy for schizophrenia. The model’s output can only be used as a reference by physicians due to its imperfect accuracy, which makes collecting MRI data directly from the scanner manufacturer problematic. Additionally, multimodal MRI input fusion is lacking, which could improve the diagnosis.

Hassanien et al. (2022) [23] proposed a method for breast tumor malignancy prediction that uses breast ultrasound sequences through a deep learning-based network that extracts radiomics features and a quality-based malignancy score pooling system. However, most previous studies have used single ultrasound images which leads to low accuracy in classification due to the presence of noise and boundary variations. To address these problems, the authors used the ConvNext architecture to capture more features from multiple frames of BUS sequences instead of SUI (single ultrasound image). Moreover, they implemented a pooling mechanism based on image quality by weighting the contribution of each frame based on its quality (brightness, blurriness) to make sure low-quality frames have less impact on the overall decision. The model was trained and tested on a dataset of 31 malignant and 28 benign cases consisting of 3911 and 5245 images respectively. It outperformed current single image-based models like MobileNetV3 with 91.66% accuracy, 93.05% precision, and a 92.33% F1 score. However, their work faced limitations such as sensitivity to image quality variation. The authors suggested that future research should focus on verifying their method using other datasets on breast cancer.

In this study [26], the authors presented ConvNeXt, a modified convolutional neural network (ConvNet) architecture, in this research to maintain the usefulness of ConvNets in comparison to Vision Transformers (ViTs) for image recognition applications. ConvNets were the industry standard for computer vision jobs until the 2020s when Vision Transformers (ViTs) were released and quickly took over the field. They addressed the issue of ConvNets losing their relevance. To bridge this gap, the authors brought a revised architectural update where a standard ResNet is modernized with the addition of design elements from ViTs. The updates include improved training techniques, larger convolutional kernels, a patchify stem, some macro design modifications, and updated activated functions. The model was trained and evaluated on extensive datasets, including ImageNet-1K, ImageNet-22K, COCO, and ADE20K, for benchmarking purposes. ConvNeXt beat the existing ViTs in object detection and segmentation tasks and scored 85.5% top-1 accuracy on ImageNet-1K when tested on image classification, object detection (using COCO), and semantic segmentation (using ADE20K). It proves that modernized ConvNets can be competitive in the market full of transformer architectures restoring its lost glory. However, the authors mentioned some limitations in areas of transfer learning in other tasks and it requires further scaling compared to vision transformers.

In this research [27], the authors developed a deep ensemble learning model called ViTCNX for the automatic detection of COVID-19 using lung CT images [1]. The problem addressed in the paper is the difficulty of distinguishing COVID-19 cases accurately from other types of pneumonia and healthy cases. Another key problem is varying image quality and limited labeled data of CT images. The authors combined the Vision Transformer(ViT) and ConvNext models to utilize the full potential of both architectures enhancing classification accuracy. The authors combined three lung CT datasets resulting in a total of 7398 CT images where 3768 CT images are of COVID-19 patients, 2383 CT images of healthy cases, and 1247 CT images of other pneumonia patients. The outputs of ViT and ConvNeXt were merged to provide a final prediction in the authors’ ensemble model, which used a soft voting method. In the binary classification of COVID-19 vs. healthy patients, the ViTCNX model obtained an accuracy of 98.21%, a recall of 0.9907, and an F1-score of 0.9855. The results show that the model outperforms other state-of-the-art architectures like DenseNet, Swin transformer. However, the limitation of this model is the increased computational costs due to the ensemble’s complexity which needs more training time and higher computational resources.

The authors developed a deep learning approach for early diagnosis of Alzheimer’s Disease (AD) using brain MRI data in order to improve classification accuracy between AD and cognitively normal (CN) individuals in this paper [21]. The paper addresses the problems of low accuracy and reliability in existing models for fluctuations in image quality of brain MRI. To overcome this, the authors proposed using tissue segmentation techniques combined with convolutional neural network models (CNN). The idea behind this is that focusing on key brain tissues could enhance the model’s feature extraction capabilities. The ADNI dataset was used where there were 435 subjects of which AD and CN groups were divided evenly and 35 CN subjects were discarded. Pre-processing steps contained brain extraction, normalization, and image registration, followed by segmentation of brain tissues us-

ing K-means clustering and Hidden Markov Random Field (HMRF) models. Then the segmented images were used to train various CNN models such as ResNet, VGG-16, and GoogleNet, comparing their performance against models trained on non-segmented images. The results showed that ResNet trained on segmented images achieved the max accuracy of 93.75% whereas, it achieved only 90.83% when trained on non-segmented images. That proves tissue-labeled images outperform unsegmented images in classification. However, the author faced limitations in training instability and aimed to reduce the issue by adding more stepping values. Also, they aim to enhance MRI images using other modalities and utilizing image segmentation which will further improve the AD vs CN classification.

In the paper [25], the writers introduce a unique way of using 3D CNN to distinguish between schizophrenia and normal controls based on resting state fMRI data. Usually other papers in this field discards a lot of important data during preprocessing but here they used the whole NIFTI files that contains valuable information about brain activity. An extraction method called "reverse phase de-ambiguity" is used which increases the signal-to-noise ratio by highlighting BOLD-related signals and reducing noise. SSPNet is designed to fully capture the complex spatial relationships hidden in the 3D phase maps. Furthermore, the model utilizes interpretability features such as saliency maps and Gradient-weighted Class Activation Mapping (Grad-CAM) and highlights relevant brain areas involved in schizophrenia. As a result, clinicians and researchers will understand the classification results better and debug accordingly. The results of this study indicate that SSPNet is efficient and superior in classification. It outperforms existing methods that rely solely on magnitude data and is better in accuracy, sensitivity and specificity. For instance, the model shows a significant improvement in classification accuracy and Area Under the Curve (AUC) metrics. This study improves diagnostic accuracy and helps us to develop the world of neuro-diagnosis using machine learning.

Across all of these papers that were reviewed, one of the major drawbacks was the lack of sufficient datasets, which limited scalability and made the results less robust and generalizable. Due to this, most of the models showed reduced accuracy when tested on independent datasets. Furthermore, computational complexity was another factor that most of these research had in common. It was noticed that the transformer and other hybrid architecture were the primary culprits. And lastly, most of these studies relied on single data modalities instead of multimodal integration. Considering all the common challenges, we will take the necessary steps to try to avoid facing them. First of all, we will gather as many different datasets as we can from all accessible sources. We will also make sure that the dataset contains diverse and quality data. Secondly, in order to improve the results, we will integrate multimodal data as inputs. Finally, we will optimize our model for speed and efficiency such that it may not delay the runtime.

Chapter 3

Methodology

3.1 ConvNeXt

ConvNeXt model is a pioneering deep learning architecture that combines the strengths of convolutional neural networks (CNNs) and cutting-edge training methods utilized in Vision Transformers (ViTs). From the beginning, the CNNs has been the go to for any kind of computer vision tasks like image classification, object detection and segmentation. In fact it is not a coincidence as the concept of sliding window to extract features of high resolution images is crucial in computer vision field. Moreover, unlike transformers CNN architectures have translational equivariance that is suitable for object detection tasks. However, with the rise of transformers, originally created for natural language processing (NLP) tasks have made its way into the field of image analysis. In 2020s, ViTs took over the ConvNets and outperformed in several image analysis tasks. Next came the ConvNeXt architecture that brought ConvNet architecture back into the competition beating Vision Transformers in many areas. The ConvNet architecture was modernized by incorporating design elements and training techniques from Vision Transformers to give the birth of ConvNext architecture [26]. A standard ResNet architecture was chosen and modernized with improvements such as a "patchify" stem, larger kernels, inverted bottleneck layers and swapping Batch Normalization with Layer Normalization (LN). Further tweaks have been made in activation functions like GELU instead of RELU and fewer activation functions. These adjustments helped to compete neck and neck with state-of-the-art ViTs like the Swin Transformer in visual recognition tasks like ImageNet-1k and COCO object detection. ConvNeXt achieves 85.5% ImageNet-1k top-1 accuracy and 87.8% ImageNet-22k top-1 accuracy which proves that it can still perform competitively or better than similarly sized Swin Transformers with a little greater throughput. The model displayed the true potential of ConvNet architecture when modernized achieve high accuracy in visual recognition tasks with very simple architecture compared to Vision Transformers.

3.2 Vision Transformer

The Vision Transformer model (ViT), is a deep learning model that is used for image segmentation and classification. At first, transformers were introduced for machine translation jobs as an alternative to recurrent and convolutional neural networks, and they have since been widely used in many NLP applications. Then for capturing long-range dependencies and connections between different parts of an image, ViTs were created by adopting the transformer architecture, as an alternative to CNNs [33]. When operating with large datasets, ViTs are highly efficient if they are sufficiently pre-trained. They break down a large image into smaller patches which is similar to the tokens used in NLP. This means that an image of resolution $B \times H$ (breadth x height) is divided into $P \times P$ size of pixels. So, each patch can be treated like tokens in the case of NLP. Each of them is divided into a size of $P \times P \times C$. In this case, the pixels are denoted as P , and C is the total number of channels for the image. Later on, they are flattened by a linear transformation and turned into a proper-sized vector known as embedding. This is a crucial stage since it ensures all the vectors have a fixed size, and can finally be fed to the transformer which works on the sequence of embeddings. Finally, these vectors are used in transformers in order to know about the connection between the small parts of the image. These transformers have layers that consist of multi-head self-attention mechanisms and feed-forward neural networks. Through the self-attention mechanism, contextual information is captured throughout the whole image. Then, a unique classification token is appended to the embedding sequence in order to facilitate classification tasks. By combining data from every patch, this token enables the model to generate an ultimate prediction about the complete image. According to research, ViTs are capable of achieving state-of-the-art results in semantic segmentation, proving their capacity to extract context and minute information from the input images [13]. However, there are certain limitations of ViTs. Sometimes ViTs require a lot of data and also have significant computational costs. With their ability to capture long-range dependencies through self-attention mechanisms, ViTs provide a potential method for image segmentation and classification. However, because of how much their performance depends on massive datasets and processing capacity, more research is necessary to improve their effectiveness and generalizability across a range of industries, including medical imaging [17]. However, the architecture of transformers cannot get the accurate idea about the 2d spatial relation between the patches as it was originally made keeping in mind only plain texts. That's why positional embeddings are used in ViTs. Through this, the transformer gets to know about each patch's position accurately. Due to this trait despite being very data-hungry, ViTs have a sense of freedom to look anywhere on the image and learn where to focus on to give accurate predictions. Unlike CNNs, ViTs can learn more adaptable, global representations of pictures. So, it has become an immensely popular choice for computer vision tasks.

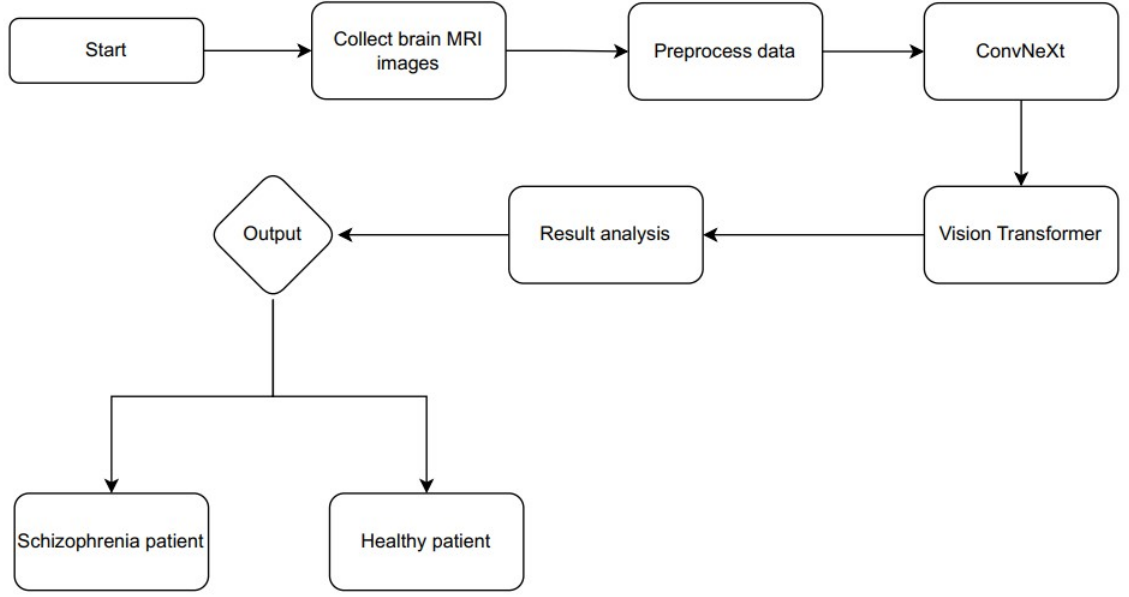


Figure 3.1: Methodology Steps

3.3 Work Plan

We propose the following steps to conduct our research:

- **Data Collection:** We obtain brain MRI images of both schizophrenia and healthy patients from publicly available datasets.
- **Data preprocessing:** In this step, we will perform several preprocessing steps on the MRI images to ensure training efficiency and enhance model performance.
- **Model design:** We use a pre-trained Vision Transformer and ConvNeXt model and fine-tune it on the brain MRI dataset. The Vision Transformer model is used for initial global feature extraction. In the second stage, we extract detailed features from the ConvNeXt model.
- **Testing and evaluation:** We test the hybrid model on an unseen test dataset and note down the accuracy, precision, recall, and F-1 score metrics.
- **Model Analysis:** We use the results and compare it with other similar models to determine its impact in real life scenarios and any limitations along with future possibilities.

3.4 COBRE Dataset

3.4.1 Dataset Collection

For our work, we are using a secondary dataset collected from the Center for Biomedical Research Excellence (COBRE) dataset [7]. The sample consists of 72 schizophrenia patients (58 males) and 74 healthy controls (51 males). The age of each group ranges from 18 to 65 years. Each subject has a 3D+t nifti file (.nii.gz) with fMRI data from the resting state. It features 150 EPI blood oxygen level-dependent volumes (BOLD) acquired in 5 mins ($TR = 2$ s, $TE = 29$ ms, $FA = 75^\circ$, 32 slices, voxel size = $3 \times 3 \times 4$ mm³, matrix size = 64×64 , $FOV = 128 \times 128$ mm²). The data set was pre-processed using the NeuroImaging Analysis Kit (NIAK).

The COBRE dataset was selected as our primary dataset for fine-tuning and evaluation purpose for our hybrid model. Moreover, COBRE is an open-access dataset available through the International Neuroimaging Data-sharing Initiative (INDI) that ensures transparency and reproducibility of our experiments. In the field of neuropsychiatric disorder detection through neuroimaging, COBRE is a widely used benchmark. There are plenty of studies where it was utilized to train deep learning models for schizophrenia detection making it an ideal choice for our study.

3.4.2 Dataset Preprocessing

The fMRI dataset went through several preprocessing processes to make sure of the data quality and consistency. Motion correction has been used here in order to account for head movement, followed by slice timing correction and intensity non-uniformity correction to align and improve the pictures. To allow cross-subject comparison, each of their median fMRI was co-registered with their T1 scan and afterwards transformed into the MNI space by applying the CIVET pipeline. Then, by merging all the spatial transforms i.e by the use of rigid-body transform, fMRI-to-T1 transform, and T1-to-stereotaxic transform, the functional volumes were resampled at 3mm isotropic resolution. Scrubbing was used to minimize high-motion volumes (frame displacement > 0.5 mm), and temporal high-pass filtering eliminated low-frequency drifts. Nuisance regression was employed to exclude motion related and non-neuronal traffic in the data. Lastly, the spatial smoothness of the dataset was obtained by a 6 mm Gaussian blur, and also was used to reduce the noise.

After acquiring the preprocessed 4D Nifti files (T, D, H, W), each volume was reduced to a 3D representation by computing the voxel-wise standard deviation across the time dimension. W, H, D represents width, height and depth respectively in voxels whereas T represents temporal dimension (number of time points or volumes). Then 0.1th percentile and 99.9th percentile values of voxel intensities were calculated and normalized to $[0,1]$. In MRI data, some voxels might have artificially high values due to scanner noise or preprocessing artifacts, this helps ignore those extreme outliers and focus on meaningful signal. After that clipping was performed out of range voxels to $[0,1]$. Because even after percentile normalization, a few values

might still fall slightly below 0 or above 1, clipping forces all voxels to stay within $[0,1]$. Each 3D volume was then resized to $64 \times 64 \times 64$ (D, H, W) using anti-aliased interpolation. Finally, each volume was converted to a $1 \times 64 \times 64 \times 64$ (C, D, H, W) PyTorch tensor making it ready for input to model. C represents channel dimension.

3.5 OpenNeuro UCLA Dataset

3.5.1 Dataset Collection

In this study we use the UCLA Consortium for Neuropsychiatric Phenomics LA5c Study which is a public resource available via OpenNeuro (dataset ID: ds000030) [11]. The dataset reports on 265 participants which includes healthy controls ($n = 130$), people diagnosed with schizophrenia ($n = 50$) bipolar disorder ($n = 49$) and adult ADHD ($n = 43$). Only control and schizophrenia groups is chosen for pretraining into model. Participants' age range was between 21 to 50 years (mean age 33.23 years; 155 men and 157 women) and they were either native English or Spanish speakers with at least eight years of formal education. Data was acquired using a 3T Siemens Trio MRI machine. Functional MRI data was collected using a T2* weighted echoplanar imaging (EPI) sequence which had a TR of 2000 ms, TE of 30 ms, flip angle of 90° , 34 axial slices, a slice thickness of 4 mm, matrix size of 64×64 and field of view of 192 mm. Structural scans were done with a T1 weighted MPRAGE sequence which had a TR of 1900 ms, TE of 2.26 ms, 176 slices, slice thickness of 1 mm, FOV of 250 mm, matrix of 256×256 . Also included were resting state and several task based fMRI sessions (i.e. BART, PAMRET, SCAP, STOPSIGNAL, TASKSWITCH, Breath Hold). Not all subjects took part in each task due to either design changes or incomplete data collection.

In our study, the UCLA LA5c dataset was used specifically for pretraining the hybrid model. Since, the dataset had a larger sample size compared to the COBRE dataset, pretraining on this dataset allowed the model to learn more general brain features like what a control or schizophrenia brain might look like before fine-tuning on COBRE dataset. This technique is called transfer learning where in the beginning a model learns from one dataset, then applies that knowledge to a different but related dataset. Without pretraining the hybrid model could likely overfit or perform poorly.

3.5.2 Dataset Preprocessing

Preparation of the UCLA LA5c dataset was done with FMRIPREP version 0.4.4 which in turn uses tools from FSL, ANTs, FreeSurfer, and AFNI. T1 weighted anatomical images were bias corrected, skull stripped and non linearly registered to MNI space using ANTs. Cortical surfaces were reconstructed with FreeSurfer that also used high quality masks from ANTs.

Functional MRI data was motion corrected with FSL's MCFLIRT, skull stripped,

and coregistered to anatomical images via boundary based registration. All transformations which included normalization to MNI space were done in one go to preserve data quality. Confound regressors were obtained from white matter and CSF signals also along with motion parameters and framewise displacement. Each subject’s data includes motion corrected and normalized BOLD images, confound regressors and quality reports. The preprocessed 4D NIFTI (D, H, W) data went through all the stages mentioned in COBRE dataset preprocessing. After the completion of those stages, $1 \times 64 \times 64 \times 64$ (C, D, H, W) PyTorch tensor is made for input to model.

3.5.3 Dataset Augmentation

To increase the generalization performance of the model and reduce overfitting caused by the small size of medical imaging sets we applied a variety of data augmentation techniques during training. We used the TorchIO library which is tailored for 3D medical imaging to do this. The augmentation pipeline was applied to each fMRI volume with variable probabilities, which in turn imitated real world anatomical variation seen in neuroimaging data.

The transformations included random flipping along the left-right, anterior-posterior, and superior-inferior axes which we applied a probability of 0.8 to introduce orientation variation. Also we applied random affine transformations which included scaling by a factor from 0.6 to 1.4 and rotation up to ± 40 degrees with a probability of 0.8. to simulate scanner and position related variations. We applied elastic deformations with 15 control points and max displacement of 20 voxels for 70% of the time ($p=0.7$) to put in non-linear anatomical distortions which modeled individual structure variation (differences in brain folding). We added Gaussian noise with standard deviation from 0 to 0.3 using probability 0.7 to simulate scanner noise. Also we applied a random bias field for 70% of the time ($p=0.7$) to model scanner related intensity nonuniformities and added gamma transformations in log range of $[-0.8, 0.8]$ with probability ($p=0.8$) for variability in brightness and contrast across scans. Finally we introduced random anisotropy 60% of the time ($p=0.6$) to represent non-isotropic voxel spacing. Here voxel dimensions are different along each axis that improves generalization with different resolutions.

3.6 ConvNeXt Architecture

The neural network which is used in our paper is a specialised version of the ConvNeXt architecture, adapted for the analysis of 3D volumetric data by the author for medical image analysis [28]. In our hybrid model, we used ConvNeXt to extract the hierarchical features. We built our ConvNeXt framework such that it can take the 3D inputs of fMRI as shown in the Figure 3.2. Using those inputs, our model extracts the spatial features from the function MRI so that it can be used later on by the ViT component of our model. Here, ConvNeXt module takes input in the shape of $([B, 1, 64, 64, 64])$ where 1 is the channel dimension and the rest are the depth, height and width dimensions of the 3D data $([B, C, D, H, W])$. The B value

represents the batch size. First of all, the 3D brain scans are sent through a hierarchical structure consisting of a stem layer which pre-process the data using a $4 \times 4 \times 4$ kernel size and a stride of 4. Afterwards a group normalization is applied to the 64 channels. As a result, the output of the 3D data is reduced by a factor of 4 which means $64/4 = 16$ and the 1 channel has been mapped to 64 channels. Our stem layer helps us to process the raw input image before it goes to the deeper parts of the network. The resulting output dimension thus changes to $([B, 64, 16, 16, 16])$. Our architecture consists of four stages after the stem layer. Each process features at progressively lower resolutions but higher semantic levels. We set our ConvNeXt blocks per stage to $([2, 2, 6, 2])$, which means stage 1, 2 and 4 will have 2 blocks, while stage 3 will have 6 ConvNeXt blocks. And our channel count per stage is set to $([64, 128, 256, 512])$ which means stage 1 will have 64 channels, stage 2 will have 128 channels, stage 3 will have 256 channels and stage 4 will have 512 channels. Furthermore downsampling layers, each with kernel size of $2 \times 2 \times 2$ with stride 2 are added in between every stage, except after the stem layer as it is handled internally, such that the input dimensions are reduced at every layer after stage 1. Stage 1 receives the input layer of $([B, 64, 16, 16, 16])$ and outputs to stage 2 the same dimension of $([B, 64, 16, 16, 16])$ due to the inactive downsample layer. Again stage 2 takes this input dimension and outputs $([B, 128, 8, 8, 8])$, which is then passed on to stage 3 that outputs $([B, 256, 4, 4, 4])$. Finally in stage 4, the final output dimension is $([B, 512, 2, 2, 2])$. Along with downsampling between stages, group normalization is done at the end to stabilize the inputs.

Our ConvNeXt block architecture, shown in the Figure 3.3 consists of several layers. It starts with a 3D depthwise convolution with a $7 \times 7 \times 7$ kernel size and with padding of 3. It applies the kernel to all channels. For example, in stage 1 there are 64 channels so it will apply 64 independent $7 \times 7 \times 7$ kernels which means one per channel. The output dimension will be similar to the input dimension. Then group normalization takes place across all channels within 1 group. After this inverted bottleneck happens, which is a vital element of the ConvNeXt architecture inspired by the transformer block. It starts by rearranging the input dimension of $([B, C, D, H, W])$ to $([B, H, W, D, C])$ which goes to pointwise convolutions and expands the channel dimension 4 times the original C value. This is followed by a GELU(Gaussian Error Linear Unit) activation function which introduces non-linearity in the model. Finally another pointwise convolution takes place that compresses the channel dimension by 4 times so that it returns to the original C value. Finally the inverted bottleneck module is completed by reverting the dimension from $([B, H, W, D, C])$ to the standard format of $([B, C, D, H, W])$. Layer scaling is applied to it channel-wise for training stability by controlling the contribution of the block's output to the residual sum. Finally to prevent vanishing gradient, the residual connection is established by adding the input to output, along with a very low dropout rate for regularization and reducing overfitting.

To conclude the whole ConvNeXt process of feature extraction, global pooling is done over the last three dimensions $((H, W, D))$. Here, input dimension of $([B, 512, 2, 2, 2])$ from stage 4 produces $([B, 512])$ as output dimensions after global pooling. Therefore, ConvNeXt output is a 512-dimensional feature vector per sample. The entire data flow is shown in a clear and concise way along with our ConvNeXt block.

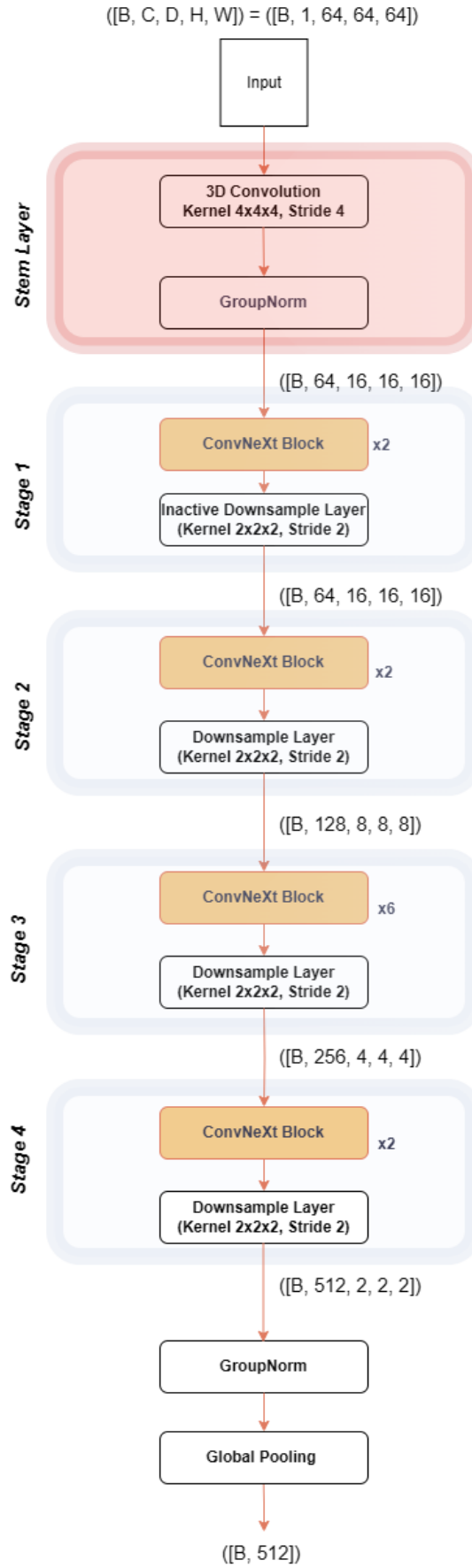


Figure 3.2: Overview of the ConvNeXt Architecture of our Hybrid Model

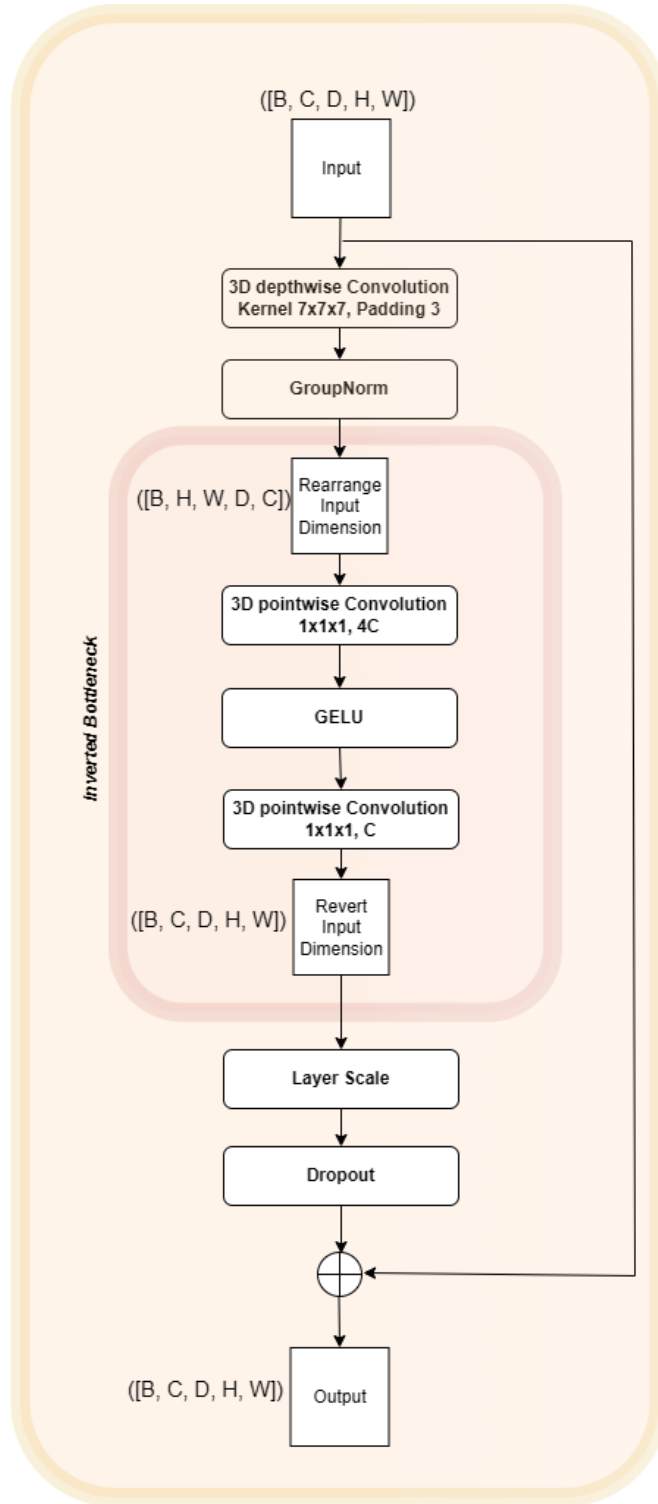


Figure 3.3: Overview of the ConvNeXt Block of the ConvNeXt Architecture

3.7 Vision Transformer Architecture

In the hybrid model, the Vision Transformer component as shown in the Figure 3.4, has been designed for capturing the long range dependencies and model global attention by leveraging the self-attention mechanisms from the 3D fMRI data. After the hierarchical features extracted by the ConvNeXt module from volumetric brain scans, ViT is used to encode relationships between the spatial patches using transformer layers. Thus, the model can make more accurate predictions about brain activity patterns that might indicate schizophrenia.

The ViT receives 3D fMRI scan of shape $([B, C, D, H, W])$, where B is the batch size, C is the input channel length and the remaining are spatial dimensions representing the depth, height and width respectively. Here the 3D input value given to the vision transformer is $[B, 1, 64, 64, 64]$. Though the volume is divided into non-overlapping patches, the preprocessing enhances input before embedding.

$$x \in R^{B \times 1 \times 64 \times 64 \times 64} \quad (3.1)$$

Patch shifting is a feature augmentation scheme that is default in Swin Transformers [19]. It thinks in generating a number of spatially shifted replicas of the input data and summing them, thereby presenting more diversity to the spatial properties and rendering the model invariant to small spatial variations. This not only helps in improving generalization to unseen data (as a model) but also helps in better local context and spatially dependent captures, which are important in interpretation of convoluted 3D brain images. Thus for preserving spatial awareness, shifted 3D views of the input have been used. To enrich the spatial context, four shifted versions of the input volume have been created which occur along depth, height and width dimensions by half the patch size increasing the channel dimension from 1 to 5 (1 original + 4 shifted). Then the shifted volume is cropped $([B, 1, 64, 64, 64])$ and concatenated with the original to augment contextual diversity resulting in tensor of shape $[B, 5, 64, 64, 64]$.

$$X_{\text{shifted}} = \text{concat}(X, S_1(X), S_2(X), S_3(X), S_4(X)) \in R^{B \times 5 \times 64 \times 64 \times 64} \quad (3.2)$$

A 3D convolutional layer is applied with kernel size and stride equal to patch size (16) used for patch embedding. Each 3D patch is assigned a coordinate-based positional encoding so that each patch location in original brain volume is known. Three coordinate channels (x, y, z) are concatenated to the input. The simple concept of CoordConv [8] circumvents a limitation of standard convolutions, which are translation-invariant and thus do not observe absolute position. The CoordConv layers make the network learn the difference between the same patterns that happen in various areas of the brain, which is crucial to the tasks like brain disorder classification, where the same signal in the different brain regions could have different clinical meanings.

$$X_{\text{coords}} = \text{concat}(X_{\text{shifted}}, x_c, y_c, z_c) \in R^{B \times 8 \times 64 \times 64 \times 64} \quad (3.3)$$

This layer embeds each patch (16 * 16 * 16 voxel cube) into a feature vector of dimension 192 (embed_dim = 192). The spatial dimension reduce from 64 to 4

($64/16 = 4$), yielding $[B, 192, 4, 4, 4]$ while the patching results in $4 * 4 * 4 = 64$ total patches.

$$P = \text{Conv3D}(X_{\text{coords}}) \in R^{B \times 192 \times 4 \times 4 \times 4} \quad (3.4)$$

The output is flattened ($[B, 192, 4, 4, 4]$ to $[B, 192, 64]$) and transposed ($[B, 192, 64]$ to $[B, 64, 192]$).

$$T = \text{reshape}_{B,64,192}(P) \quad (3.5)$$

A learnable class token with 192 features (the embedding dimension, which is the size of the vector representing the token) denoted as

$$\mathbf{c} \in R^{B \times 1 \times 192}; \text{ class token dimension} \quad (3.6)$$

This acts as a dynamic “summary agent” is expanded to $[B, 1, 192]$ for a batch of B volumes and is concatenated with the patch embeddings acting as a dynamic “summary agent” is expanded to $[B, 1, 192]$ for a batch of B volumes and is concatenated with the patch embeddings forming the resulting tensor,

$$Z_0 = [c; T] \in R^{B \times 65 \times 192}; \text{ resulting tensor after concatenation} \quad (3.7)$$

Positional embeddings are added to encode spatial information which are generated from a grid of patch coordinates mapped to 192 dimensions through a linear layer.

$$Z_0 \leftarrow Z_0 + E_{\text{pos}} \text{ where } E_{\text{pos}} \in R^{1 \times 65 \times 192} \quad (3.8)$$

For regularization a dropout layer is applied.

For the integration of local (ConvNeXt) and global (ViT) features, the ConvNeXt feature vector of each item of 512 dimensional feature is,

$$\mathbf{F}_{\text{convnext}} \in R^{B \times 512} \quad (3.9)$$

This is projected down to ViT’s patch embeddings of 192 dimensional space using a linear layer (like a simple matrix multiplication):

$$\mathbf{F}_{\text{proj}} = W_f \cdot \mathbf{F}_{\text{convnext}}^\top \in R^{B \times 192} \quad (3.10)$$

Thus the result is a new tensor with shape $[B, 192]$. Now, as the ViT has already created the patch embeddings of 64 patches, the expanded ConvNeXt features are then added element-wise to the patch embeddings and injects localized ConvNeXt hierarchical information into each transformer token enhancing ViT’s understanding of the fMRI data.

$$\mathbf{f}_{\text{exp}} = \mathbf{f}_{\text{proj}} \cdot \text{expand}(-1, 64, -1) \in R^{B \times 64 \times 192} \quad (3.11)$$

Updated patch sequence with localized context,

$$\mathbf{P}' = \mathbf{P} + \mathbf{f}_{\text{exp}} \quad (3.12)$$

$$Z_0 = [\mathbf{c}; \mathbf{P}'] \in R^{B \times 65 \times 192} \quad (3.13)$$

The transformer consists of 4 repeated blocks as shown in Figure 3.5, each containing locality self-attention and Feed-Forward Network (MLP). In locality self-attention, input is projected to queries, values via linear layer, split into 6 heads, each of dimension 32 ($192/6$).

$$Q = Z_i W_Q, \quad K = Z_i W_K, \quad V = Z_i W_V \quad (3.14)$$

Attention score,

$$A = \frac{QK}{\sqrt{d_h}}, \quad d_h = 32 \quad (3.15)$$

A locality mask is used to prevent self-attending and a learnable temperature for scaling attention sharpness,

$$\alpha = \text{softmax} \left(\frac{A}{\tau} \right), \quad \tau \in R^{6 \times 1 \times 1} \quad (3.16)$$

Attention output,

$$\tilde{V} = \alpha V \quad (3.17)$$

After reshaping, the final attention output,

$$A_l = \tilde{V}^{\text{flat}} W_O \quad (3.18)$$

In Feed-Forward Network, two linear layers with a GELU activation in between where the first linear layer expands the 192 dimensional vector to 768, GELU activation helps the model to learn complex patterns, the second layer projects the 768 dimensional vector to 192. Then, like the attention step, the output is added to attention and MLP blocks for stability.

$$U_l = Z_l + \text{LN}(A_l) \quad (3.19)$$

$$F_l = W_2 \cdot \text{GELU}(W_1 \cdot U_l), \quad W_1 \in R^{192 \times 768}, W_2 \in R^{768 \times 192} \quad (3.20)$$

$$Z_{l+1} = U_l + \text{LN}(F_l) \quad (3.21)$$

The final transformer output is [B, 65, 192] or,

$$Z_4 \in R^{B \times 65 \times 192} \quad (3.22)$$

Then by extracting the class token representation of last transformer layer, the final output of ViT is obtained and is normalized,

$$\mathbf{y} = \text{LayerNorm}(Z_4[:, 0]) \in R^{B \times 192} \quad (3.23)$$

The resulting 192 dimensional vector is subsequently passed through a classification head for performing final binary decision.

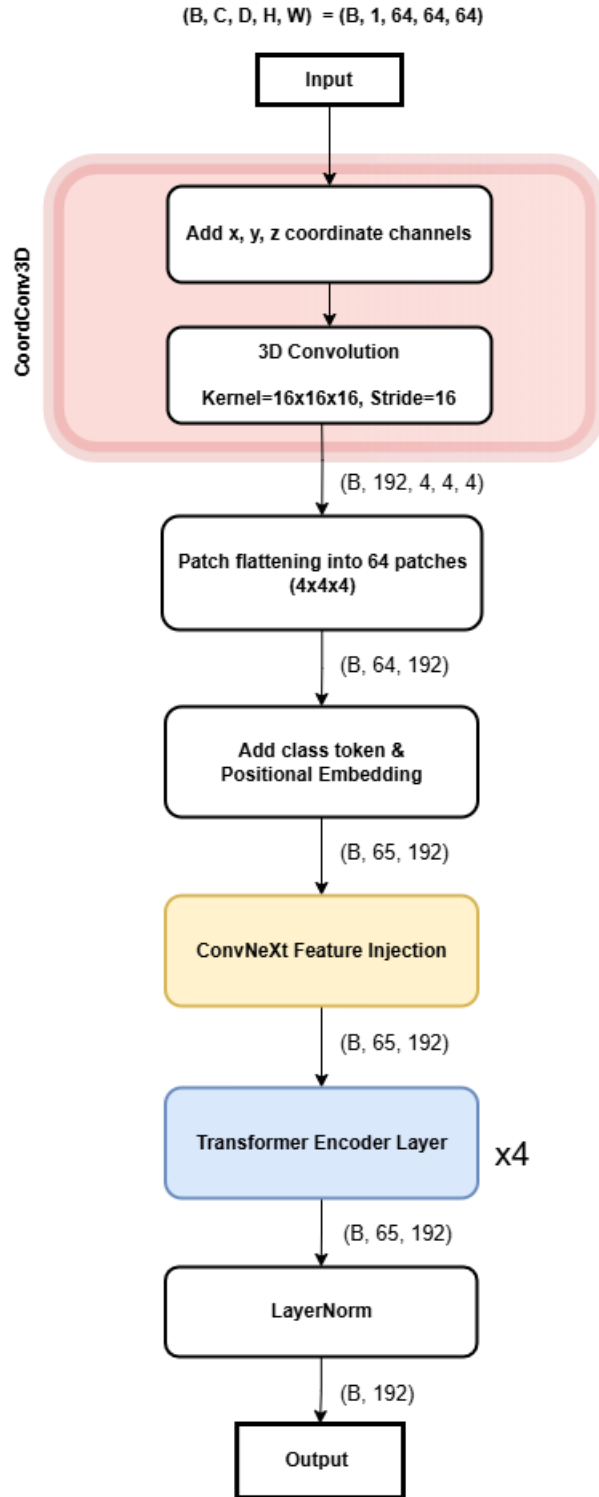


Figure 3.4: Overview of the ViT Architecture of our Hybrid Model

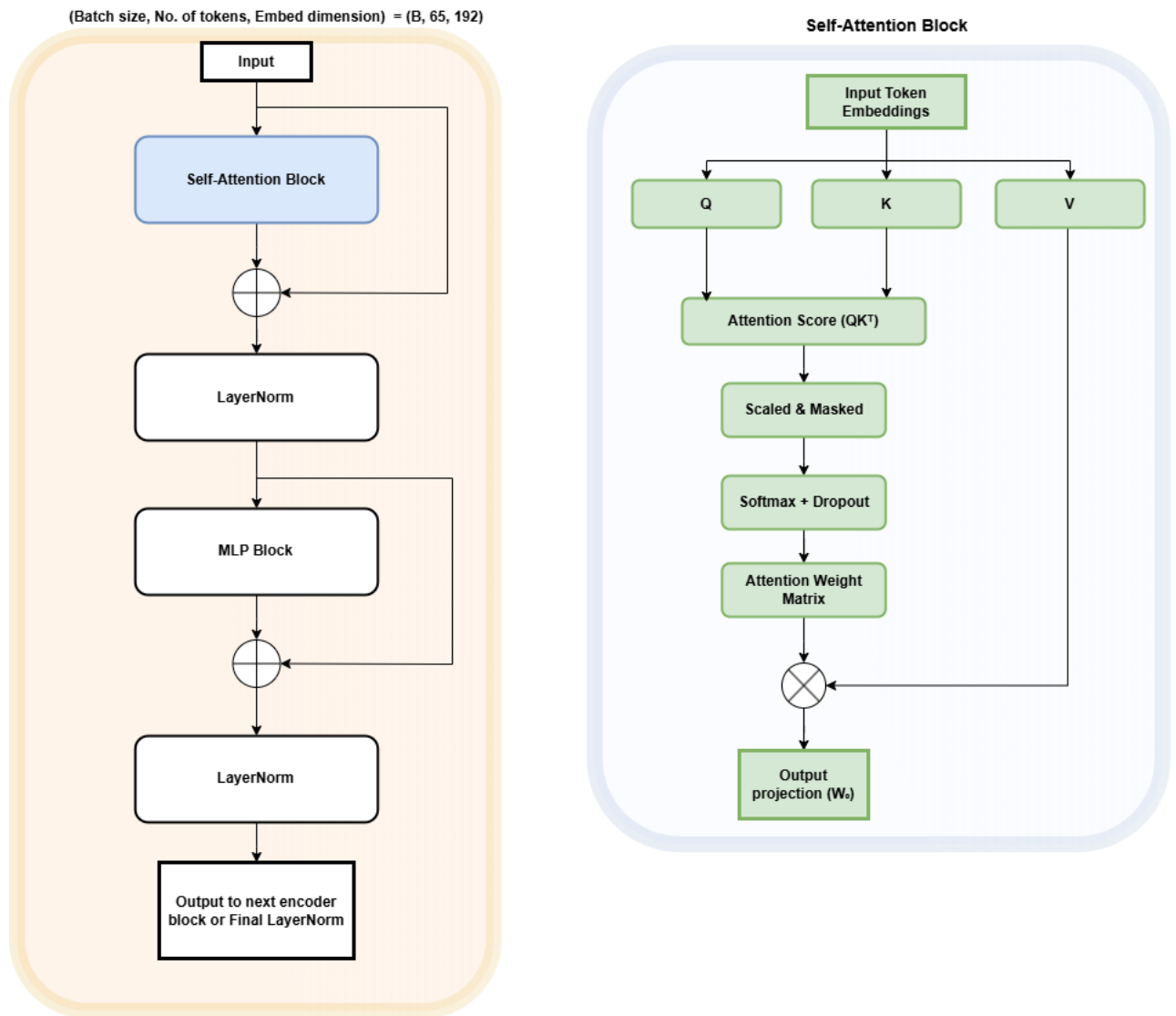


Figure 3.5: Overview of ViT Transformer Encoder and Self-Attention blocks

3.8 Hybrid Architecture

The proposed hybrid architecture as shown in the Figure 3.6 combines the strengths of both ConvNeXt3D and Vision Transformer (ViT) models through the integration of convolutional and attention based representations during forward pass. It allows the model to capture both low level local features and high level global relationships for effective schizophrenia classification.

3.8.1 ConvNeXt Output and Feature Projection

Firstly a 3D fMRI input of shape $[B, 1, 64, 64, 64]$ is given to the ConvNeXt3D feature extractor which in turn produces a rich spatial feature map that is then mean pooled across the spatial dimensions and after that is normalized. Out of this a feature vector of shape $[B, 512]$ where B is the batch size and 512 is the final embedding size, is obtained from ConvNeXt.

To make the transformer encoder compatible with this output the 512 dimensional ConvNeXt output is passed through a linear projection layer that reduces the dimension of the vector to match the ViT embedding size (192 dimension). The projected vector is reshaped to $[B, 1, 192]$ and then broadcasted out to every patch token in the ViT excluding the [CLS] token which in turn the patch is enriched with high level context generated by ConvNeXt output.

3.8.2 Transformer Encoding and Output Extraction

After merging the Vision Transformer (ViT) patch embeddings with the ConvNeXt features, the embedded token sequence is passed as input to four transformer encoder layers. These layers help the model to learn better on how different parts of the brain relate to each other using self-attention.

Each transformer layer consists of two major components: locality self-attention and a Feed-Forward Network (FFN). Both components are connected by residual connections and normalized using Layer Normalization (LayerNorm) thus forming a deep residual transformer block. The input to the Locality Self-Attention block is a tensor of shape $[B, 65, 192]$ where B = Batch size, $65 = 64$ patch tokens + 1 class token [CLS] and $192 =$ embedding size. The attention block makes three copies of the input embedding: query (Q), key (K) and value (V) matrices and then splits it across six attention heads. These projections are then reshaped for multi-head attention. A diagonal mask is applied to stop a token from focusing on itself and each head carries a learnable temperature value to adjust the sharpness of the softmax distribution. Afterwards, the model computes attention scores by dot product between the query and key matrices. The masked and scaled attention scores are then passed through a softmax activation function, converting them into normalized attention weights. These weights are used to compute a weighted sum over the value vectors. It aggregates contextual information from the other tokens.

Dropout is applied to the attention weights to prevent overfitting. Finally, the output from all heads is concatenated and projected back to the original embedding dimension of 192. Then this output is added to the input tensor via a residual connection. Next comes the Layer Normalization which helps to maintain stable gradients.

Following the attention block, the output is passed through a feed-forward neural network (FFN). The FFN consists of two linear layers with a non-linear GELU activation function in the middle. The first linear layer expands the embedding size from 192 to 768 (a 4x expansion). The second layer reduces it back to 192 dimensions. Dropout of probability of 0.4 is applied after the activation for regularization. Like in attention block the output of the FFN is added to its input through a second residual connection. Layer Normalization is again applied after the addition. Each transformer layer preserves the original input shape throughout its processing. After passing through all four transformer layers, the token sequence maintains the shape $[B, 65, 192]$. Here each token has been progressively enriched with global contextual information.

3.8.3 Class Token Extraction

Once the final transformer layer outputs the processed token sequence the model isolates the class token for downstream classification. The class token is the first token in the sequence and it is intended to serve as a summary representation of the entire input volume. By design, this token interacts with all patch tokens during the attention phases. The class token output is of shape $[B, 192]$ which gathers information about the relationships between different brain regions.

3.8.4 Final MLP Classifier and Output Generation

The extracted class token is passed into a simple neural network, multi-layer perceptron (MLP) to make the final prediction. The classifier consists of the following layers: linear projection layer, GELU activation, dropout Layer and final linear layer.

In the linear projection layer, the class token of shape $[B, 192]$ is passed through a fully connected layer that projects it to a 256-dimensional vector $[B, 256]$. The output of the linear layer is passed through a GELU activation function. It introduces non-linearity and smoother gradient propagation. A dropout layer with a probability of 0.4 is applied for regularization which reduces the risk of overfitting by randomly deactivating neurons during training. The resulting tensor of shape $[B, 256]$ is projected down to $[B, 2]$ that gives the final output logits for binary classification (schizophrenia vs. healthy control). During evaluation, softmax turns the output into probabilities. During training, the raw logits go into a weighted cross-entropy loss to handle class imbalance. The hybrid model learns to minimize this loss with gradients propagated back through the entire architecture.

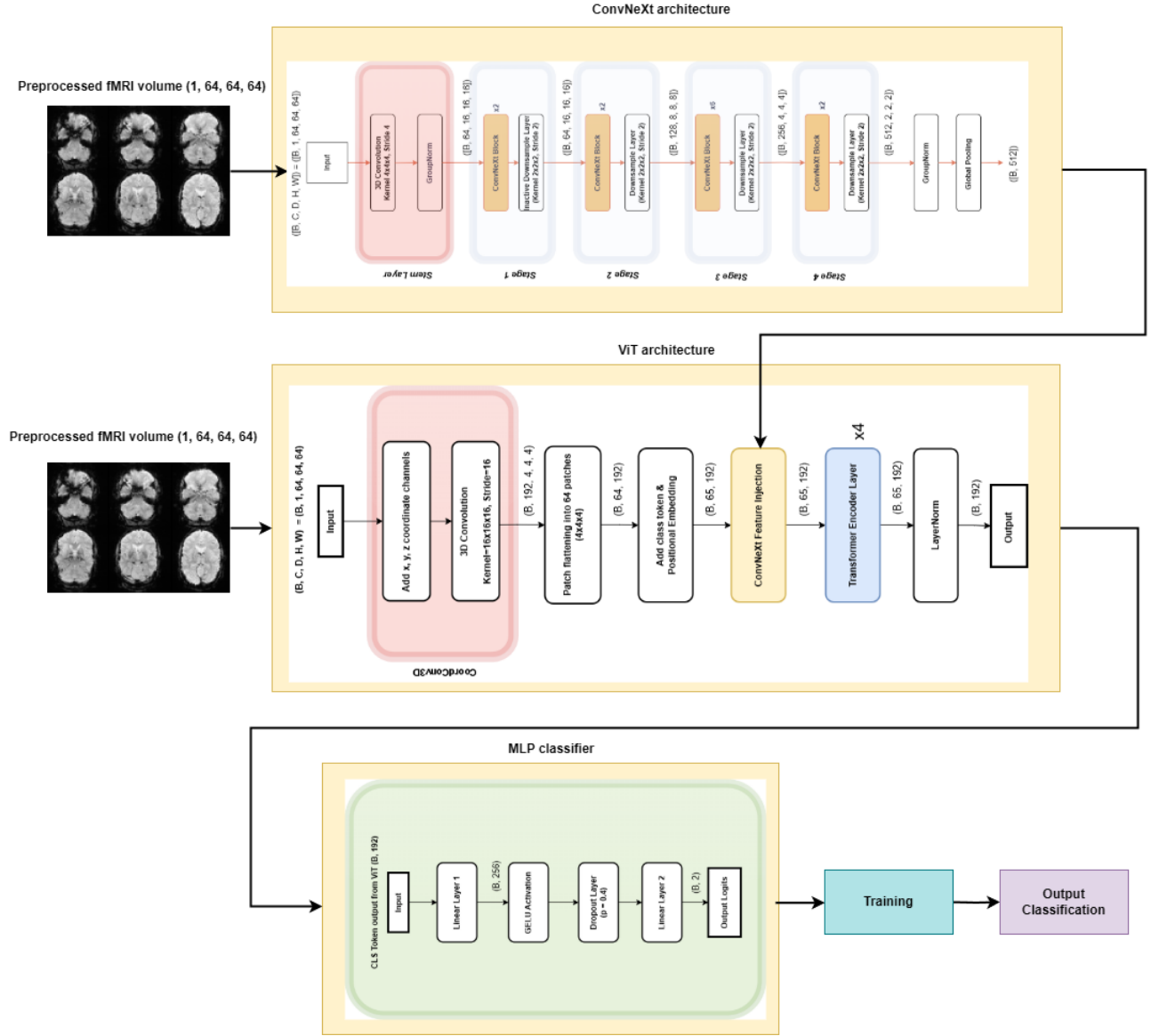


Figure 3.6: Overview of our proposed Hybrid Architecture for Classification

Chapter 4

Result and Analysis

4.1 Training and Evaluation Protocol

Training was executed on a CUDA-enabled single machine running on Windows 10 Pro with 12th Gen Intel(R) Core(TM) i5-12400F with Nvidia GeForce GTX 4080 Super and 16GB RAM device using the PyTorch framework.

For our Hybrid model, we used AdamW optimizer with an initial learning rate of 1×10^{-5} and weight decay of 1×10^{-4} to penalize large weights so that overfitting is avoided. To change the learning rate over time, a learning rate scheduler is implemented following the cosine annealing where the minimum learning rate is 1×10^{-6} . This ensures that the model starts learning more at the start and gradually slows down at the end of the training process. AdamW improves the generalization of deep neural networks by decoupling weight decay from the gradient-based parameter updates. In each training step, the model parameters θ are updated in the following manner:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla L_t \quad (4.1)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) (\nabla L_t)^2 \quad (4.2)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (4.3)$$

$$\theta_t = \theta_{t-1} - \eta_t \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_{t-1} \right) \quad (4.4)$$

Here, η_t is the learning rate at epoch t , λ is the weight decay coefficient, and β_1, β_2 are momentum terms.

To adaptively control the learning rate, we applied the following Cosine Annealing schedule:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left[1 + \cos \left(\frac{t}{T_{\max}} \pi \right) \right] \quad (4.5)$$

where η_{\max} and η_{\min} are the initial and final learning rates, and T_{\max} is the total number of epochs. This scheduler allows for an aggressive initial learning phase which is followed by gradual refinement as the learning rate is decreased. This is particularly effective for deep models that are trained on small and imbalanced datasets. It helps reduce the risk of premature convergence and improves final model generalization while reducing overfitting.

Gradient clipping is used to avoid gradient explosion during backpropagation. Furthermore, mixed precision training is enabled from our PyTorch library to reduce memory usage and speed up computation time. A custom Cross-entropy loss was used as the loss function to compute the difference between the predicted probabilities and the actual class labels, with weights adjusted to 0.493 for the schizophrenia class and 0.507 for the control class to address class imbalance. The formula for the weighted cross-entropy loss is as follows:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \alpha_{y_i} \cdot \log \left(\frac{\exp(x_{i,y_i})}{\sum_{j=1}^C \exp(x_{i,j})} \right) \quad (4.6)$$

Here N is the batch size and C is the number of classes. $x_{i,j}$ is the un-normalized logit output of the model on sample i and class j . y_i is the true class of sample i while α_{y_i} is the weight of that class. Directly in our code, we use the a PyTorch funtion that takes raw logits and performs the softmax internally and also performs the negative log-likelihood outside. The purpose of using this weighted version of the loss is because the schizophrenia dataset has an imbalance in classes. The class weights are calculated using the inverse frequency of a given class. This balance will not make the model biased toward the majority class and it will learn to be more concerned with the misclassifications done on the minority class and this will make the model more competent in identifying schizophrenia even with skewed training data.

The dataset was divided into five folds for cross-validation (K-fold), ensuring that each fold was used as a test set once while the others formed the training set (K-1). With a batch size of 8 for GPU efficiency, training is configured by a maximum epoch value of 100 per fold with early stopping of patience 50 and delta (δ_{y_i}) value of 0.005 to maximize AUC metric. Both the stratified K-fold cross-validation and early stopping was implemented to stop the model from overfitting. The following algorithm was utilized for our work.

1. Initialize `best_score = None`, `wait = 0`
2. For each epoch:
 - (a) If `current_score` improves `best_score` by at least δ :

- i. Update `best_score`
 - ii. Reset `wait` to 0
- (b) Else:
 - i. Increment `wait`
- (c) If `wait` \geq `patience`:
 - i. Stop training

To convert model probabilities into binary class predictions after training, we systematically searched for the optimal decision threshold. We evaluated a range of thresholds and selected the value that maximized the F1 score on the validation set while using accuracy as a tiebreaker when needed. In this way, we ensured that the classification boundary is best suited to the imbalanced nature of our data. The following pseudocode was used for this process.

1. Define a list of candidate thresholds (e.g., from 0.30 to 0.60, increasing by 0.01).
2. Initialize `best_threshold` to 0.5 and `best_F1` & `best_accuracy` to 0.0.
3. For each `threshold` in the list of candidate thresholds:
 - (a) Convert predicted probabilities to binary predictions using:
 $\text{prediction} = 1 \text{ if } \text{probability} \geq \text{threshold}, \text{ else } 0.$
 - (b) Compute the F1 score and accuracy for these predictions.
 - (c) If the F1 score is greater than `best_F1`, **or**
 if the F1 score equals `best_F1` and accuracy is greater than `best_accuracy`:
 - i. Update `best_threshold` to the current `threshold`.
 - ii. Update `best_F1` to the current F1 score.
 - iii. Update `best_accuracy` to the current accuracy.
4. After checking all thresholds, return the `best_threshold`, `best_F1`, and `best_accuracy`.

Prior to running this model on the COBRE dataset, pre-training was achieved using the UCLA dataset to enhance generalization on our small dataset. The code along with the results [42] are included for better understanding.

4.2 Evaluation Metrics

Accuracy It measures the overall proportion of cases the model got right among both patients and controls. It gives a quick sense of general performance, but can be misleading if classes are imbalanced (more healthy controls than patients):

$$\frac{TP + TN}{TP + TN + FP + FN}$$

F₁ Score It measures the Harmonic mean of precision and recall and balances the imbalance between false positives and false negatives. It is especially useful when you care equally about avoiding missed diagnoses (FN) and false alarms (FP):

$$2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC-ROC This is the area under the curve plotting true positive rate (recall) against false positive rate (1 – specificity) making the decision threshold. It is the probability that a randomly chosen patient will score higher than a randomly chosen control. Because it considers all thresholds, it provides a threshold-independent measure of how well the model separates the two classes and the closer it is closer to 1.0 the better it is.

TP True Positive

FP False Positive

FN False Negative

TN True Negative

Precision Proportion of true positives: $\frac{TP}{TP + FP}$

Recall/Sensitivity Proportion of actual positives: $\frac{TP}{TP + FN}$

Specificity It measures the amount of actual negatives (healthy patients) that the model correctly identifies as negative. It matters because of the low false-alarm rate. High specificity means the model rarely flags healthy patients as having schizophrenia. It helps to catch as many true patients as possible and avoid mislabeling healthy people: $\frac{TN}{TN + FP}$

Confusion Matrix 2×2 breakdown of TP, FP, FN, TN counts. Here the higher the numbers at TP and TN, the better the model is.

4.3 Hybrid Model Analysis

The hybrid model consisting of the combination of Convnext and Vision Transformer(ViT) models ran for 100 epochs per fold of total 5 folds, but it was completed by 30(approx.) epochs on every fold. The best performance was observed in Fold 2 with AUC=0.99, Accuracy=96.55 %, $F_1=0.97$, Precision=0.94, and Recall=1.00. The worst performance occurred in Fold 4, where despite perfect precision=1.00 the model missed 28.6 % of true positives that is Recall=0.7143 (71.43%), yielding AUC=0.91, Accuracy=86.2 %, and $F_1=0.83$. The training and validation losses fell from 0.75(approx.) at epoch 1 to a stable plateau around 0.70(approx.) by epoch 30, with only small fluctuations (± 0.05) thereafter and no gap between training and validation curves.

Table 4.1: Hybrid model performance by fold

Fold	Accuracy	AUC-ROC	F_1 -score	Precision	Recall
1	96.67 %	0.98	0.97	0.94	1.00
2 (Best)	96.55 %	0.99	0.97	0.94	1.00
3	89.66 %	0.99	0.91	0.82	1.00
4 (Worst)	86.21 %	0.91	0.83	1.00	0.71
5	86.21 %	0.93	0.88	0.78	1.00
Average	91.06 %	0.96	0.91	0.89	0.94

This table 4.1 lists accuracy, AUC-ROC, F_1 score, precision, and recall for each fold. Fold 2 attains the best scores (96.55 % acc, 0.99 AUC, 0.97 F_1 , 0.94 precision, 1.00 recall), while fold 4 is the weakest. The five-fold averages (91.06 % acc, 0.96 AUC, 0.91 F_1 , 0.89 precision, 0.94 recall) highlight overall robustness.

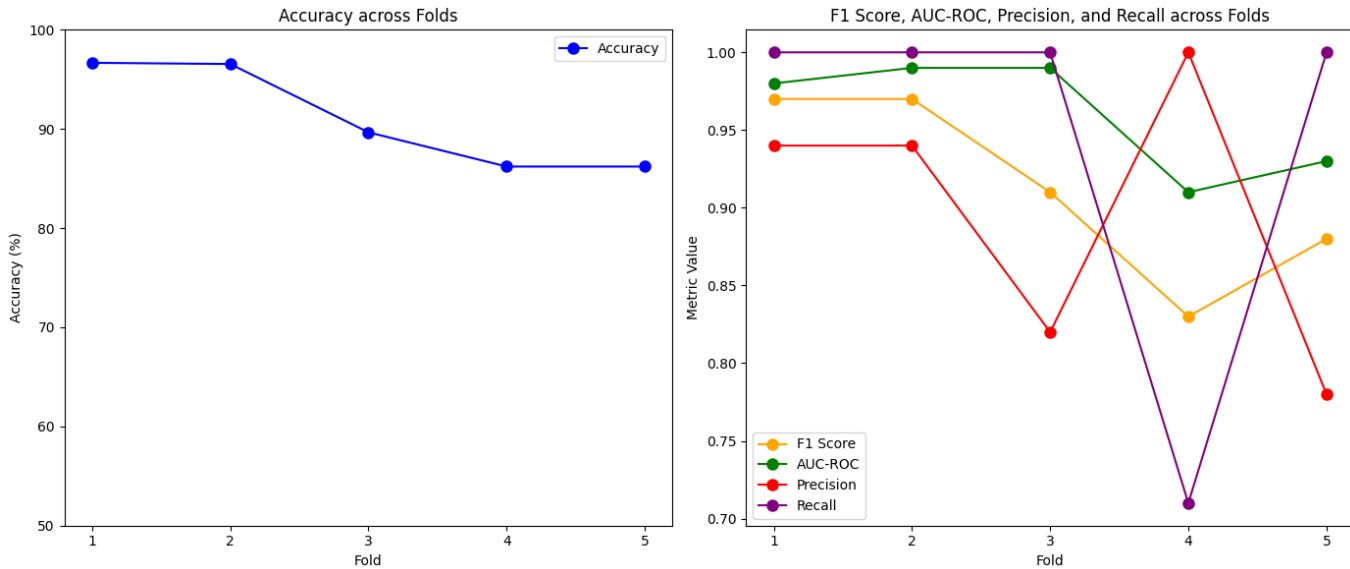


Figure 4.1: Validation Metrics across 5 folds.

Figure 4.1 shows the AUC-ROC, accuracy, F_1 score, precision, and recall computed on the validation split for each of the 5 cross-validation folds. It illustrates that all metrics rise quickly in early epochs and plateau by fold's end, with fold 2 achieving the highest overall performance.

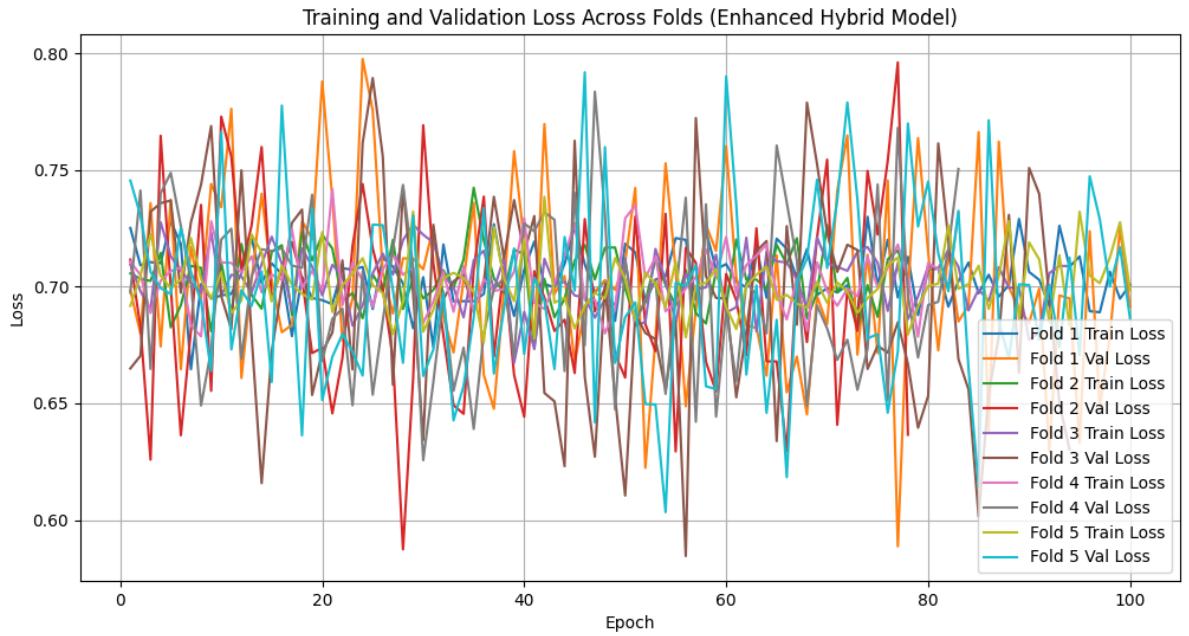


Figure 4.2: Training vs. validation loss curves.

Plots loss on both training and validation sets across epochs for all 5 folds. The close overlap of the two curves and their rapid descent to a stable minimum (0.70) by epoch 30 indicate good generalization and convergence without overfitting.

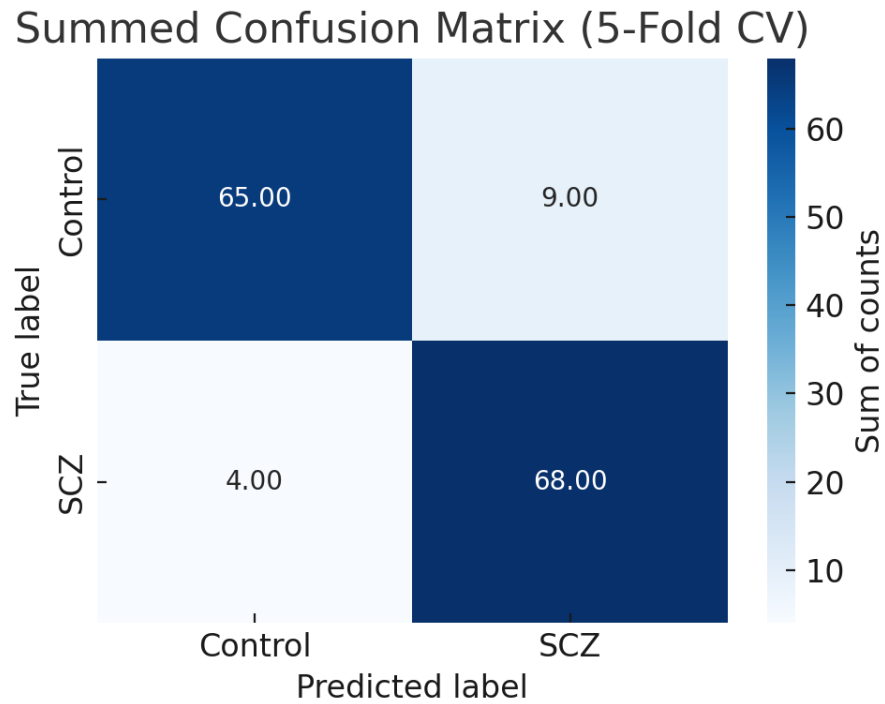


Figure 4.3: Confusion matrix summary across 5 folds.

The sum confusion matrix (TN = 65.0, FP = 9.0, FN = 4.0, TP = 68.0) in this figure 4.3 corresponds to sensitivity = 94.5 % and specificity = 88.0 %, confirming both robust disease detection and low false-positive rates on COBRE dataset under 5-fold cross-validation.

4.4 Comparing Model Analysis

3D-ResNet-18 extends the classic 2D ResNet-18 to volumetric data by replacing all 2D convolutions, batch norms, and pooling operations with their 3D counterparts. The network is built from “residual blocks,” each containing two $3\times3\times3$ convolutions and a shortcut connection that sums the block’s input with its output. These identity shortcuts alleviate vanishing gradients, enabling training of deeper architectures. The stem begins with a $7\times7\times7$ convolution and max-pool, followed by four stages of residual blocks with increasing channel widths ($64\rightarrow512$). A final global average pooling collapses spatial and temporal dimensions before a fully connected layer outputs class logits.

3D-Xception adapts the Xception architecture’s depthwise separable convolutions to 3D volumes, decomposing a standard $3\times3\times3$ convolution into a depthwise convolution (one spatial filter per channel) followed by a pointwise $1\times1\times1$ convolution for channel mixing. This factorization dramatically reduces parameters and computational cost while preserving representational power. The network stacks multiple “entry,” “middle,” and “exit” modules, each with separable convolutions and residual connections to maintain gradient flow. Strided depthwise layers reduce spatial-temporal resolution. Batch normalization and ReLU activations follow each convolution. A global pooling and classifier head complete the model.

ViT3D-B16 brings Vision Transformer principles to 3D: the input volume is divided into non-overlapping $16\times16\times16$ patches, each flattened and linearly projected into a D-dimensional embedding. Positional embeddings encode each patch’s 3D location. The sequence of patch embeddings passes through L transformer encoder layers, each comprising multi-head self-attention and feed-forward blocks with layer normalization and residual connections. Self-attention lets each patch aggregate information from all others, capturing global context. The special “class” token prepended to the sequence summarizes the volume for classification. After L layers, the class token is fed to an MLP head to predict schizophrenia vs. control.

The generic **3D-CNN** is a straightforward volumetric convolutional architecture that stacks several blocks of 3D convolution (e.g. $3\times3\times3$ kernels), ReLU activation, and 3D max-pooling to gradually reduce spatial-temporal dimensions. Early layers learn low-level features like edges and textures; deeper layers capture increasingly complex volumetric patterns. Batch normalization may be inserted to stabilize training. The convolutional backbone ends with a global average pooling that collapses the feature maps to a vector, which is fed into one or more fully connected layers with dropout for regularization, culminating in a sigmoid or softmax output for binary classification.

3D-Inception extends Google’s Inception modules to volumetric data by executing parallel convolutional filters of multiple 3D kernel sizes (e.g. $1\times1\times1$, $3\times3\times3$, $5\times5\times5$) alongside a max-pool branch within each module. The outputs are concatenated along the channel axis, allowing the network to capture features at different scales simultaneously. Bottleneck $1\times1\times1$ convolutions before expensive spatial convolutions reduce channel dimensionality, saving computation. The network alter-

nates Inception modules with grid-size reductions (via strides or pooling). A final average pooling and fully connected layer yield the classification.

3D-DenseNet-121 adapts DenseNet’s dense connectivity to 3D, organizing layers into four dense blocks separated by transition layers. Within each dense block, every layer receives as input the concatenation of all preceding feature maps, promoting feature reuse and alleviating vanishing gradients. Each layer comprises a bottleneck $1\times1\times1$ convolution followed by a $3\times3\times3$ convolution, with ReLU and batch normalization. Transition layers use $1\times1\times1$ convolutions and $2\times2\times2$ average pooling to reduce feature-map size and number. After the final dense block, global average pooling flattens the features, and a linear classifier produces the output.

3D-SwinY brings hierarchical Swin Transformer blocks to volumetric data. The volume is first partitioned into small 3D patches, linearly projected, and shaped into a sequence. Within each Swin block, self-attention is computed over local non-overlapping 3D windows, making computation linear in volume size. Blocks alternate between regular and “shifted” window partitions to allow cross-window connections. A patch-merging layer between stages halves spatial-temporal resolution and doubles feature channels, forming a hierarchical pyramid. Layer norms, MLPs, and residual connections wrap each attention block. A final global pooling and MLP head handle classification.

3D-ConvNeXt modernizes convolutional networks inspired by Vision Transformers, adapted to 3D. It replaces standard convolutions with large-kernel (e.g. $7\times7\times7$) depthwise convolutions, followed by pointwise $1\times1\times1$ layers, inverted bottlenecks, layer normalization, and GELU activations within each ConvNeXt block. The architecture is divided into four stages, each with repeated ConvNeXt blocks and occasional downsampling (strided depthwise convs). Skip connections add the block’s input to its output, preserving residual learning. After the final stage, a global average pooling and dense layer classify the volume. This design blends convolutional inductive biases with some Transformer-style normalization and activation choices.

4.5 Performance Summary Table

Table 4.2: Performance Summary of All Models

Model	Accuracy (%)	F ₁ Score	AUC-ROC	Precision	Recall
3D-ResNet-18	69.3	0.72	0.83	0.652	0.806
3D-Xception	68.1	0.66	0.69	0.697	0.639
ViT3D-B16	54.8	0.58	0.62	0.556	0.417
3D-CNN	64.4	0.55	0.64	0.727	0.444
3D-Inception	64.3	0.52	0.74	0.652	0.806
3D-DenseNet-121	58.8	0.29	0.58	0.750	0.250
3D-SwinY	52.6	0.46	0.49	0.525	0.431
3D-ConvNeXt	52.0	0.45	0.66	0.554	0.500
Hybrid (Ours)	91.06	0.91	0.96	0.883	0.944

Tabulates each model’s overall accuracy, F₁ score, AUC-ROC, precision, and recall. The hybrid clearly surpasses the baselines (e.g. 91.06 % vs. best baseline 69.3 % accuracy).

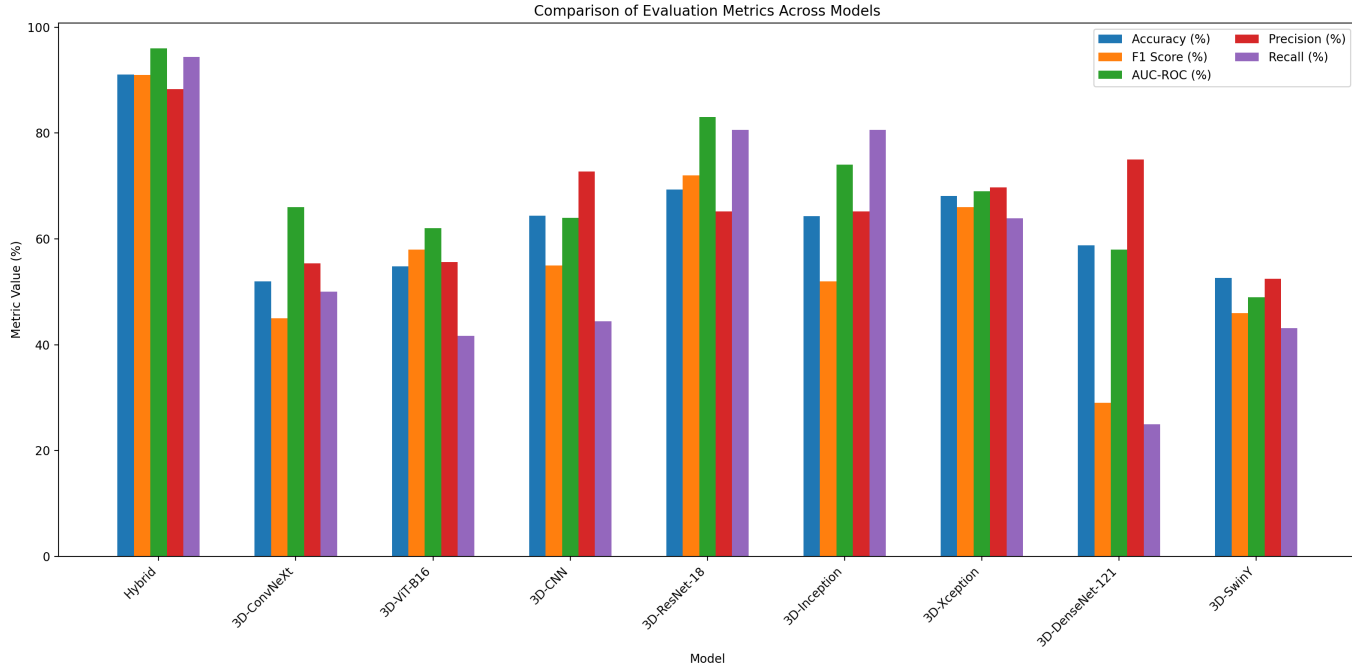


Figure 4.4: Validation Metrics of all models.

A grouped bar chart comparing accuracy, F₁ score, AUC-ROC, precision, and recall across all nine architectures (8 baselines plus the hybrid). It visually demonstrates the hybrid’s lead on every metric.

4.6 Confusion Matrix Analysis

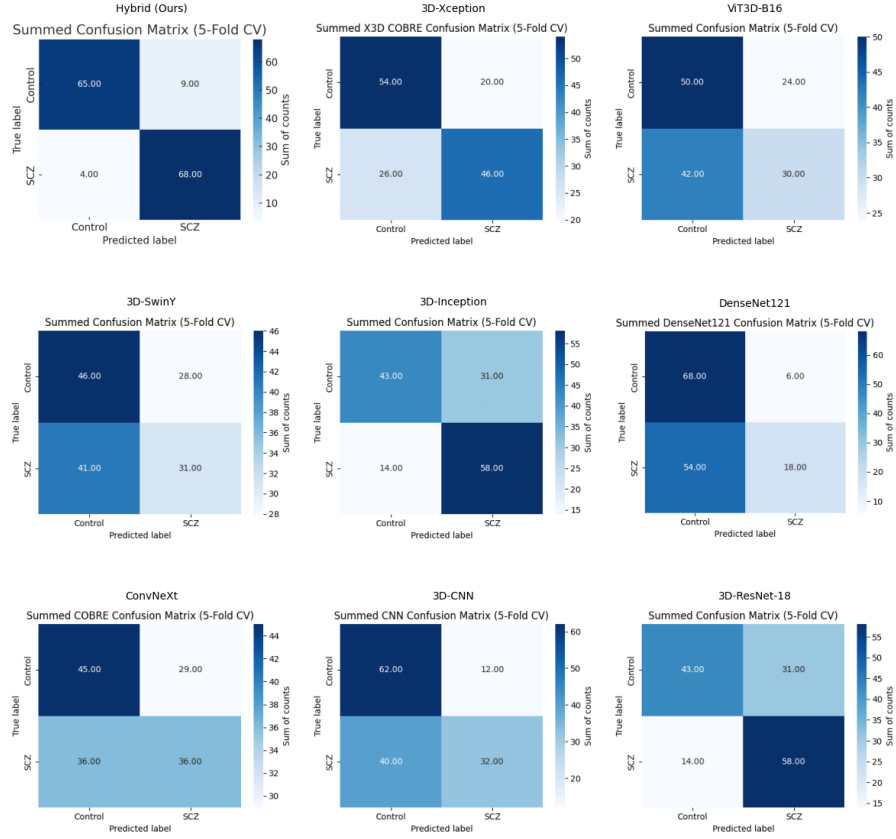


Figure 4.5: Confusion Matrix of all Models.

Nine small 2x2 heatmaps (one per model), each showing its summed TP, FP, FN, and TN. This grid highlights how few false negatives/positives the hybrid registers compared to the others.

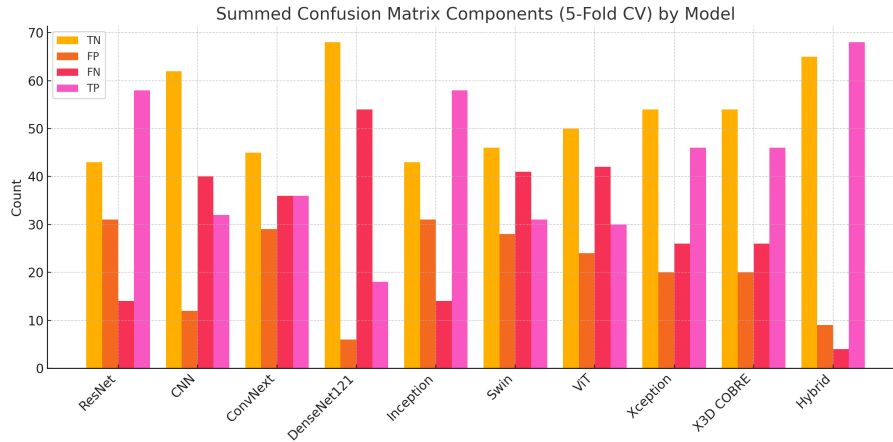


Figure 4.6: Confusion Matrix Graph of all models.

Here we can see that our hybrid model has the highest TN (65) and TP (68)

which outperforms the rest of the 8 models making it the best model for detecting schizophrenia.

Table 4.3: Confusion & derived metrics

Model	TN	FP	FN	TP	Sensitivity	Specificity
ResNet-18	43	31	14	58	0.806	0.581
Xception	54	20	26	46	0.639	0.730
ViT3D-B16	50	24	42	30	0.417	0.676
3D-CNN	62	12	40	32	0.444	0.838
Inception	43	31	14	58	0.806	0.581
DenseNet-121	68	6	54	18	0.250	0.919
SwinY	46	28	41	31	0.431	0.622
ConvNeXt	45	29	36	36	0.500	0.608
Hybrid	65	9	4	68	0.945	0.878

Lists for each model its TN, FP, FN, TP, sensitivity and specificity. The hybrid’s sensitivity (94.5 %) and specificity (88.0 %) are among the highest, reflecting its balanced performance.

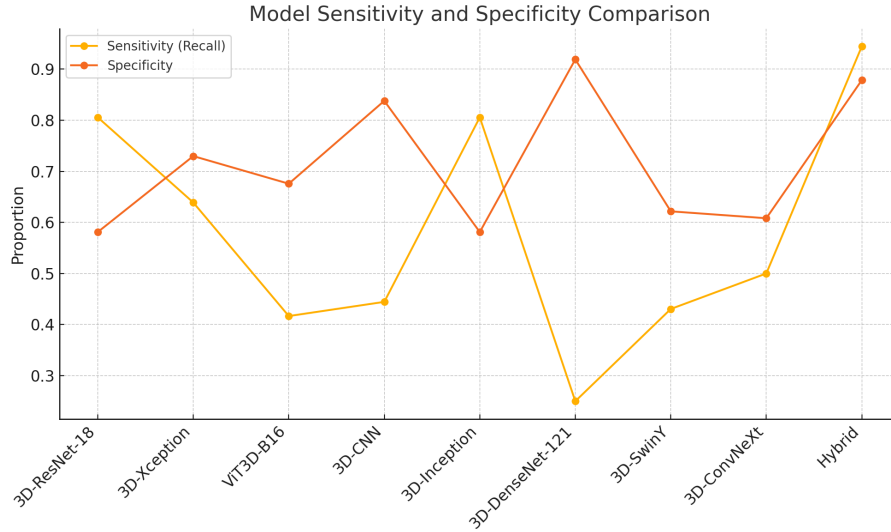


Figure 4.7: Model Sensitivity And Specificity Comparison.

Plots each model’s sensitivity versus specificity as a scatter plot demonstrating the hybrid’s position in the top-right (high on both axes), unlike most baselines which trade one for the other.

False Negatives (FN): Hybrid FN=4 is lowest, dramatically reducing missed patients compared with ResNet/Inception (14 FN) which were the closest.

False Positives (FP): Hybrid FP=9 is small, lowering mislabelled healthy cases. Although 3D-DenseNet-121 has FP=6 which is better than our hybrid model, the hybrid model’s overall confusion matrix result puts it as the better model.

Sensitivity Gains: Hybrid’s recall of 94.5 %(approx.) outperforms all baselines followed by the next best ResNet/Inception at 80.6 %(approx.).

Specificity Gains: Hybrid’s specificity of 87.8 %(approx.) is second only to DenseNet at 91.9 %(approx.), but DenseNet’s sensitivity is just 25 % which is the lowest.

Balanced Performance: Unlike baselines that trade off sensitivity vs. specificity, the hybrid achieves both high sensitivity and high specificity, and low FN and FP.

4.7 Best Epochs per Fold

Table 4.4: Run time of all models

Model	Time
Hybrid	10 h
ResNet-18	5 h
Xception	4 h
ViT3D-B16	3 h
3D-CNN	4 h
Inception	4 h
DenseNet-121	4 h
SwinY	5 h
ConvNeXt	2 h

The table lists the runtime of all the models we used in comparison with our hybrid model. All training parameters were kept the same where it was possible. In case of changes, the models were kept as similar as possible. From the table, we see that the hybrid model has the highest runtime of 10 hours while the second highest model, which is ResNet-18, has a runtime of 5 hours. The rest of the models have a runtime of around 2-4 hours.

From this information, we can say that the hybrid model had the most runtime. This is because it is a dual backbone model of ConvNeXt and Vision Transformer (ViT) which works in sequence one after the other. The 3D convolutions that we use are already really resource intensive, more so than the simple 2D convolutions. Besides, transformer models are well known for their inefficient computational costs. The fusion of these two models increases the total number of parameters and the complexity of the data flow. Thus it significantly increases memory usage and training time. Furthermore, almost all the models which were used as comparisons are lightweight models. Each of those models processes the input data using a single network architecture. Unlike the hybrid model, those models do not need to integrate outputs of one model to another one. As a result, it reduces the computational cost per training step.

Even though the hybrid model has a higher computational cost, it is worth considering due to the really high score of evaluation metrics. Furthermore, our hybrid model offers the unique idea of using both local and global knowledge for sensitive medical classification tasks, which the other models don't utilize. In the medical sector, the outcome has greater value than computational costs as lives may depend on it. In our case, it is more important to diagnose schizophrenia than it is to reduce computational costs.

4.8 Comparative Discussion

Hybrid vs 3D-ResNet-18: Across all five metrics, the hybrid model outperforms 3D-ResNet-18 by a large margin. Accuracy jumps from 69.3 % to 91.1 %, F -score from 0.72 to 0.91, and AUC-ROC from 0.83 to 0.96. Precision and recall both increase by roughly 25 % and 14 %, respectively, indicating that the hybrid not only makes fewer false alarms but also captures substantially more true schizophrenia cases. [4.8]

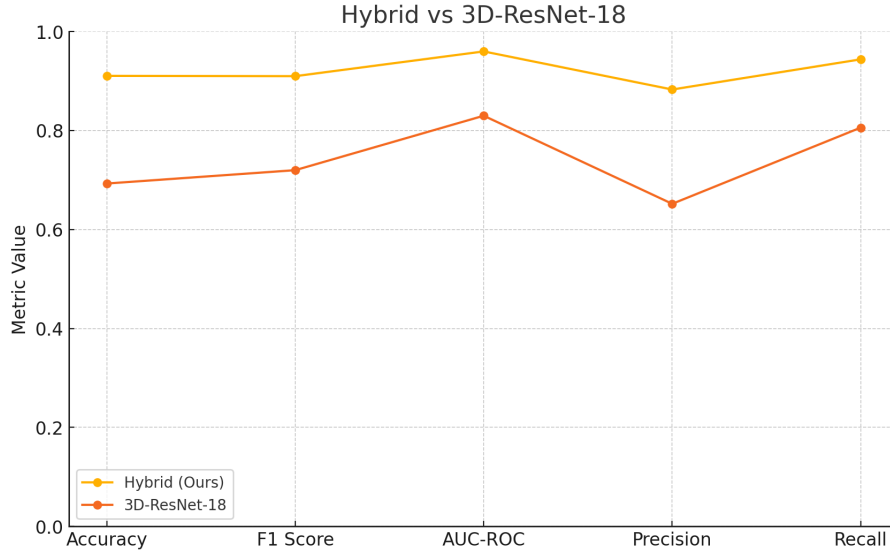


Figure 4.8: Hybrid Vs 3D-ResNet-18

Hybrid vs 3D-Xception: The hybrid’s accuracy (91.1 %) exceeds 3D-Xception’s 68.1 % by 23 %, while F -score climbs from 0.66 to 0.91. AUC improves from 0.69 to 0.96, and recall nearly doubles (0.64→0.94). Even though 3D-Xception has slightly higher precision (0.70 vs. 0.88), its low recall undermines sensitivity whereas the hybrid balances both.[4.9]

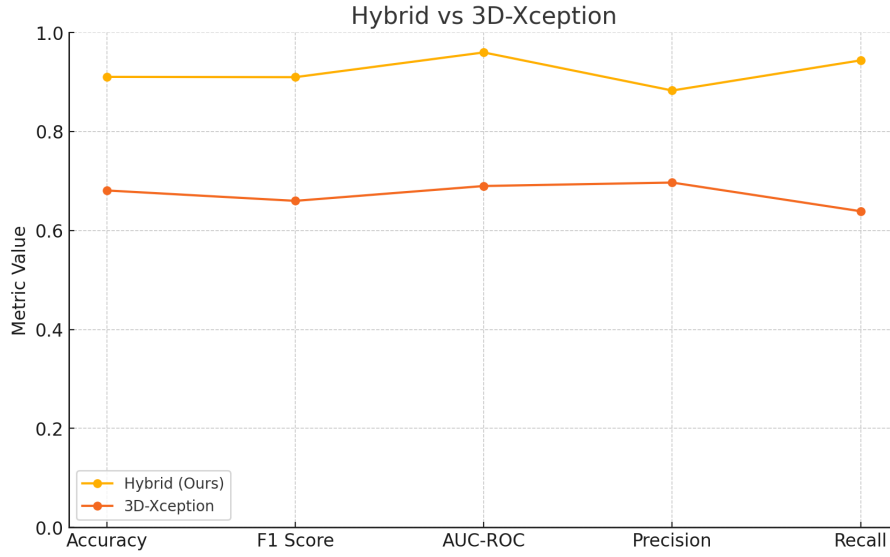


Figure 4.9: Hybrid Vs 3D-Xception

Hybrid vs 3D-ViT-B16: 3D-ViT-B16 struggles with limited data, reaching only 54.8 % accuracy and 0.62 AUC. In contrast, the hybrid achieves 91.1 % and 0.96, a gain of 36 % and 0.34, respectively. Its recall soars from 0.42 to 0.94, and F -score from 0.58 to 0.91 highlighting the hybrid’s superior ability to distinguish patients from controls. [4.10]

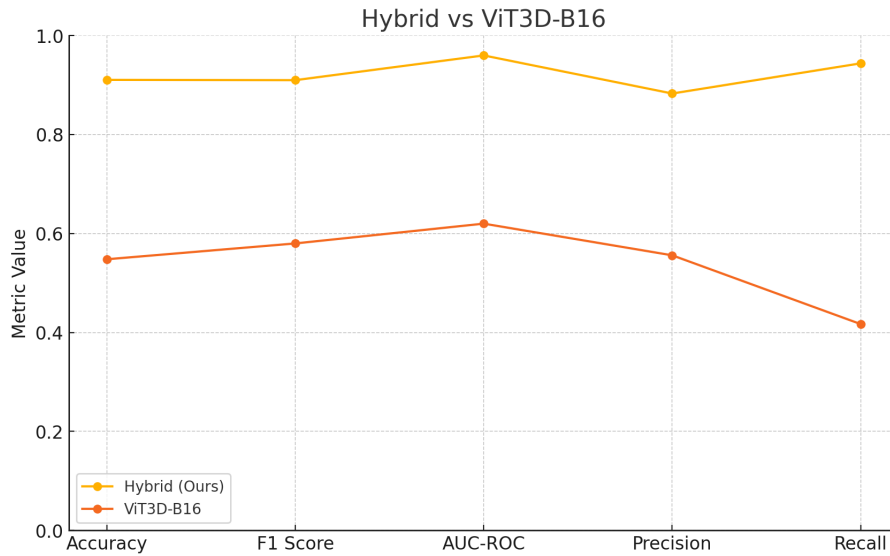


Figure 4.10: Hybrid Vs 3D-ViT-B16

Hybrid vs 3D-CNN A vanilla 3D-CNN records a modest 64.4 % accuracy and 0.64 AUC. The hybrid surpasses it by 27 % (approx.) on accuracy and 0.32 on AUC. While the 3D-CNN’s precision is comparable (0.73 vs. 0.88), its recall lags at 0.44 (vs. 0.94), meaning the CNN misses more than half of true positives, which the hybrid captures. [4.11]

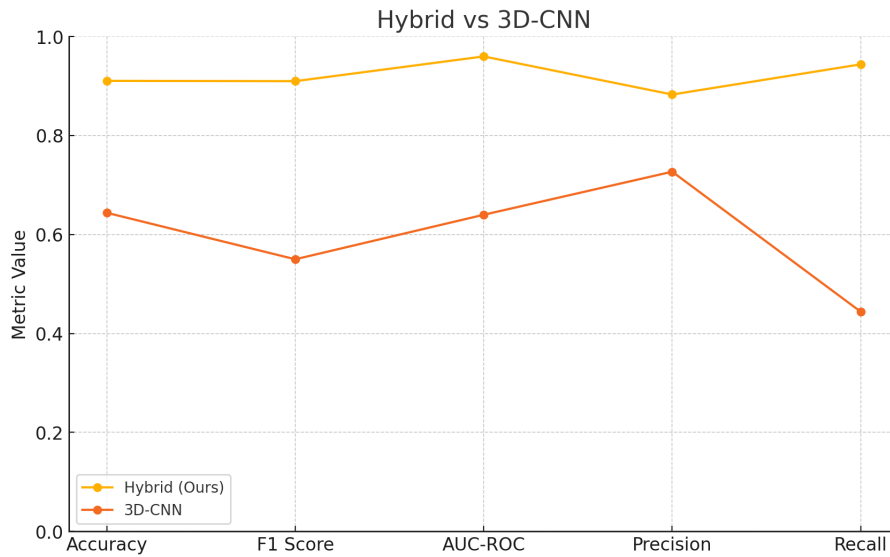


Figure 4.11: Hybrid Vs 3D-CNN

Hybrid vs 3D-Inception 3D-Inception posts 64.3 % accuracy and 0.74 AUC, with moderate F -score (0.52) and low recall (0.81). The hybrid elevates all metrics to 91.1 % Acc, 0.96 AUC-ROC, 0.91 F -score, and 0.94 recall. Its balanced precision/recall (0.88/0.94) contrasts sharply with Inception’s trade-offs, demonstrating consistently robust detection across classes. [4.12]

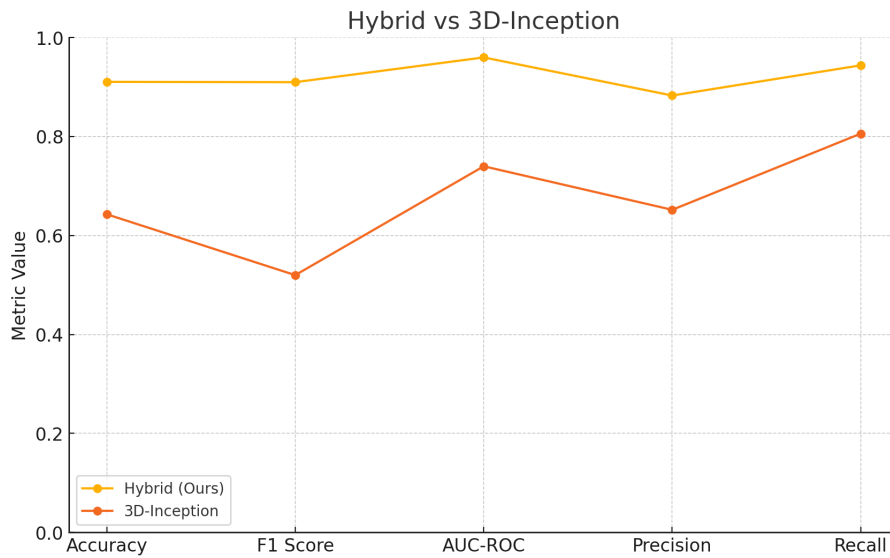


Figure 4.12: Hybrid Vs 3D-Inception

Hybrid vs 3D-DenseNet-121 DenseNet-121 shows severe underperformance (58.8 % Acc, 0.29 F -score, 0.58 AUC-ROC, 0.25 recall). The hybrid corrects nearly all of these, boosting F -score by 0.62 and recall by 0.69, and raising accuracy by 32 %. The contrast underscores the hybrid’s resistance to overfitting and its reliability to identify patients. [4.13]

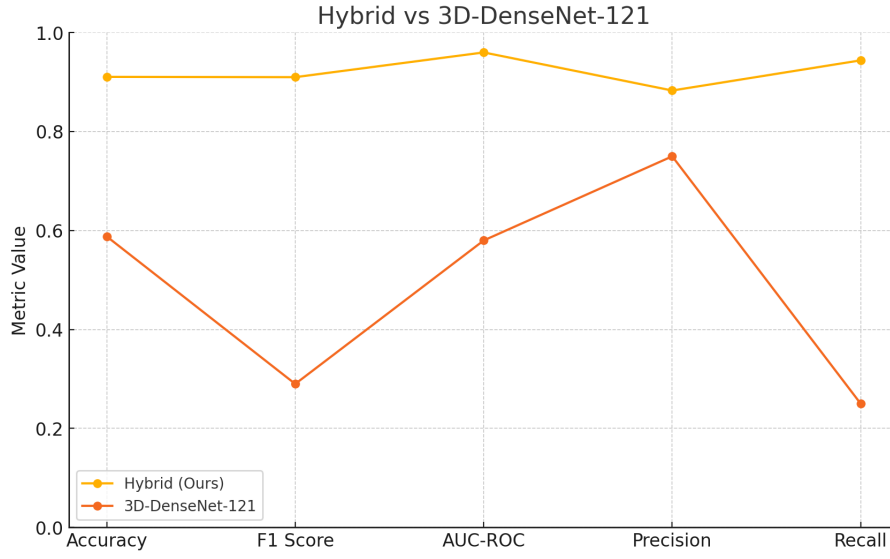


Figure 4.13: Hybrid Vs 3D-DenseNet-121

Hybrid vs 3D-SwinY 3D-SwinY plateaus at 52.6 % accuracy and 0.49 AUC-ROC, with only 0.43 recall. The hybrid doubles SwinY sensitivity (94 % vs. 43 %) and lifts accuracy by 39 %(approx.). Its AUC advantage of 0.47 more shows that the hybrid’s decision boundary is far more discriminative across thresholds. [4.14]

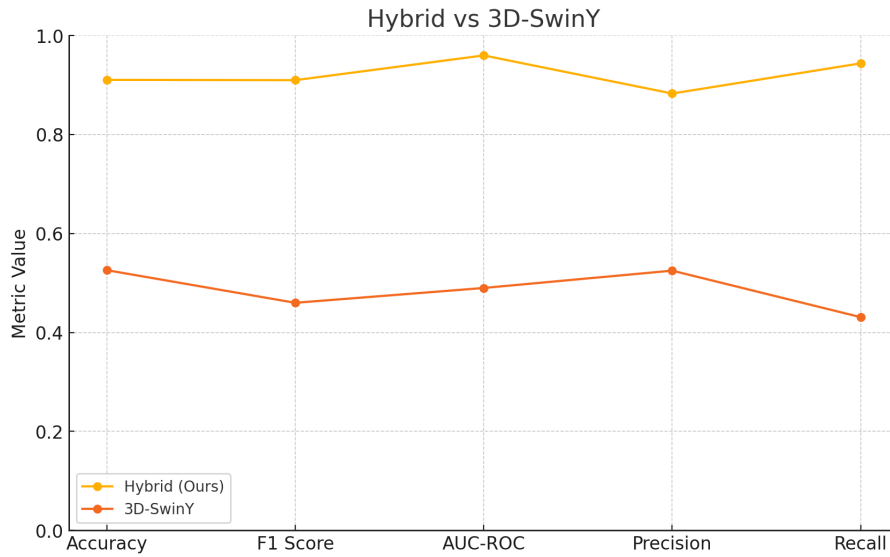


Figure 4.14: Hybrid Vs 3D-SwinY

Hybrid vs 3D-ConvNeXt ConvNeXt alone peaks at 52.0 % Acc and 0.66 AUC. The hybrid surpasses it by 39 % and 0.30 respectively, while recall improves from 0.50 to 0.94. This gap confirms that adding parallel residual and attention streams transforms ConvNeXt’s isolated strengths into a consistently high-performing model. [4.15]

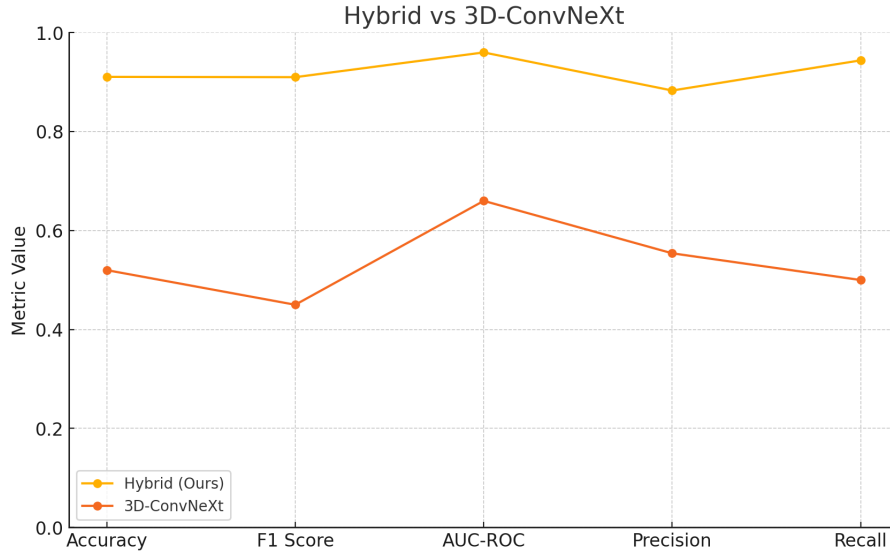


Figure 4.15: Hybrid Vs 3D-ConvNeXt

Baseline Weaknesses: 3D-ConvNeXt & 3D-SwinY underperform on small cohorts, plateauing at 52 % (approx.) Acc. 3D-ViT-B16 & 3D-DenseNet-121 overspecialize, showing near-random or zero recall in some folds. 3D-Inception & 3D-Xception suffer fold-to-fold variability (down to 48 % Acc). 3D-ResNet-18 and 3D-CNN achieve moderate AUC but lower sensitivity or specificity.

Final Verdict: Single backbone models exhibit large inter-fold swings in both metrics and best epochs whereas the hybrid’s tight band reflects robust generalization on COBRE’s limited sample amount. Our Hybrid model consisting of ConvNext and Vision Transformer (ViT) wins by fusing residual, depthwise, inception, and self-attention pathways. This model captures fine-texture, mid-level, and global context features while avoiding the single-backbone problems of under-fitting or over-fitting.

Chapter 5

Limitations and Future Work

Limitations: Our thesis has faced some limitations which obstructed us throughout the task. Firstly the main problem being the small dataset size of 146 fMRIs (72 Schizophrenia vs. 74 controls) which caused our Vision Transformer (ViT) model to not learn effectively since ViT needs lots of data to learn in depth and to give optimal performance. Second, the increased computational and memory demands of the dual-backbone architecture may constrain deployment in resource-limited clinical settings. Finally, we were not able to test our model on a live clinical environment to confirm our model’s functionality, that is we couldn’t confirm the model’s performance in a practical setting or if the model is truly good at identifying schizophrenia or not.

Future Work: For our future work, we would like to explore lightweight hybrid variants to reduce the computational load without reducing the model’s performance, expand to larger and more diverse fMRI repositories to test and to get more accurate results which will enhance our model’s performance, and to use richer explanation methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) to bridge the gap between algorithmic predictions and neuro scientific insight which will help the clinicians to understand which parts of the brain are showing schizophrenia and thus helping them make a more accurate decision. We will use ECG, PET or sMRI along with fMRI to get more in-depth information of brain connectivity. We would also like to run our model on datasets containing multiple categories such as age, gender, etc. to predict more robustly. By addressing these areas, we hope to enhance the model’s generalization, explainability, and clinical applicability.

Chapter 6

Conclusion

This thesis presents the use of a hybrid deep-learning framework that combines the local feature-extraction strengths of ConvNeXt with the global context modeling of Vision Transformers (ViT) to detect schizophrenia from resting-state fMRI scans. Across five-fold cross-validation on the COBRE dataset, the hybrid model achieved an average accuracy of 91.06%, F_1 score of 0.91, and AUC-ROC of 0.96 which substantially outperforming eight single-backbone baselines (3D-ResNet-18, 3D-Xception, 3D-ViT-B16, 3D-CNN, 3D-ConvNeXt, 3D-DenseNet-121, 3D-Inception, 3D-Swin). The hybrid model’s summed confusion matrix ($TN = 65$, $FP = 9$, $FN = 4$, $TP = 68$) gives a sensitivity of 94.5% and specificity of 88.0% demonstrating both robust patient detection and low false alarms. The hybrid model converged consistently by epoch 30 in every fold, featuring its stability and fast learning despite COBRE’s limited sample size of 146 fMRIs.

In conclusion, our Hybrid model consisting of dual-backbones works nicely by fusing residual, depthwise convolutions and self-attention pathways. This model captures fine-texture, mid-level, and global context features while avoiding the single-backbone problems of under-fitting or over-fitting. The findings speak for themselves, with consistent performance and flexibility across datasets and a significant advantage over traditional models. Although an opportunity for further development remains there, it can be a great tool for Schizophrenia detection.

Bibliography

- [1] H. Stuart and J. Arboleda-Flórez, “Community attitudes toward people with schizophrenia,” *The Canadian Journal of Psychiatry*, vol. 46, no. 3, pp. 245–252, Apr. 2001, ISSN: 1497-0015. DOI: 10.1177/070674370104600304. [Online]. Available: <http://dx.doi.org/10.1177/070674370104600304>.
- [2] G. Katti, S. A. Ara, and A. Shireen, “Magnetic resonance imaging (mri)—a review,” *International journal of dental clinics*, vol. 3, no. 1, pp. 65–70, 2011.
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- [4] P. R. Desai, K. A. Lawson, J. C. Barner, and K. L. Rascati, “Estimating the direct and indirect costs for community-dwelling patients with schizophrenia: Schizophrenia-related costs for community-dwellers,” *Journal of Pharmaceutical Health Services Research*, vol. 4, no. 4, pp. 187–194, Jul. 2013, ISSN: 1759-8885. DOI: 10.1111/jphs.12027. [Online]. Available: <http://dx.doi.org/10.1111/jphs.12027>.
- [5] M. Lee, C. Smyser, and J. Shimony, “Resting-state fmri: A review of methods and clinical applications,” *American Journal of Neuroradiology*, vol. 34, no. 10, pp. 1866–1872, 2013, ISSN: 0195-6108. DOI: 10.3174/ajnr.A3263. eprint: <https://www.ajnr.org/content/34/10/1866.full.pdf>. [Online]. Available: <https://www.ajnr.org/content/34/10/1866>.
- [6] T. M. Laursen, M. Nordentoft, and P. B. Mortensen, “Excess early mortality in schizophrenia,” *Annual Review of Clinical Psychology*, vol. 10, no. 1, pp. 425–448, Mar. 2014, ISSN: 1548-5951. DOI: 10.1146/annurev-clinpsy-032813-153657. [Online]. Available: <http://dx.doi.org/10.1146/annurev-clinpsy-032813-153657>.
- [7] P. Bellec, “COBRE preprocessed with NIAK 0.12.4,” Jan. 2015. DOI: 10.6084/m9.figshare.1160600.v15. [Online]. Available: https://figshare.com/articles/dataset/COBRE_preprocessed_with_NIAK_0_12_4/1160600.
- [8] R. Liu, J. Lehman, P. Molino, *et al.*, *An intriguing failing of convolutional neural networks and the coordconv solution*, 2018. arXiv: 1807.03247 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/1807.03247>.

- [9] O. Sankoh, S. Sevalie, and M. Weston, “Mental health in africa,” *The Lancet Global Health*, vol. 6, no. 9, e954–e955, Sep. 2018, ISSN: 2214-109X. DOI: 10.1016/s2214-109x(18)30303-6. [Online]. Available: [http://dx.doi.org/10.1016/s2214-109x\(18\)30303-6](http://dx.doi.org/10.1016/s2214-109x(18)30303-6).
- [10] B. F. Marghalani and M. Arif, “Automatic classification of brain tumor and alzheimer’s disease in MRI,” en, *Procedia Comput. Sci.*, vol. 163, pp. 78–84, 2019.
- [11] R. Bilder, R. Poldrack, T. Cannon, *et al.*, “*ucla consortium for neuropsychiatric phenomics la5c study*”, OpenNeuro, 2020. DOI: 10.18112/openneuro.ds000030.v1.0.0.
- [12] Z. Chen, T. Yan, E. Wang, *et al.*, “Detecting abnormal brain regions in schizophrenia using structural MRI via machine learning,” en, *Comput. Intell. Neurosci.*, vol. 2020, p. 6 405 930, Apr. 2020.
- [13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” Oct. 2020. DOI: 10.48550/arXiv.2010.11929.
- [14] G. Gifford, R. McCutcheon, and P. McGuire, “Neuroimaging studies in people at clinical high risk for psychosis,” in *Risk Factors for Psychosis*. Elsevier, 2020, pp. 167–182, ISBN: 9780128132012. DOI: 10.1016/b978-0-12-813201-2.00009-0. [Online]. Available: <http://dx.doi.org/10.1016/B978-0-12-813201-2.00009-0>.
- [15] J. Oh, B.-L. Oh, K.-U. Lee, J.-H. Chae, and K. Yun, “Identifying schizophrenia using structural MRI with a deep learning algorithm,” en, *Front. Psychiatry*, vol. 11, p. 16, Feb. 2020.
- [16] S. A. Ajagbe, K. A. Amuda, M. A. Oladipupo, O. F. Afe, and K. I. Okesola, “Multi-classification of alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches,” *Int. J. Adv. Comput. Res.*, vol. 11, no. 53, pp. 51–60, Mar. 2021.
- [17] J. Chen, Y. Lu, Q. Yu, *et al.*, “Transunet: Transformers make strong encoders for medical image segmentation,” Feb. 2021. DOI: 10.48550/arXiv.2102.04306.
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2010.11929>.
- [19] Z. Liu, Y. Lin, Y. Cao, *et al.*, *Swin transformer: Hierarchical vision transformer using shifted windows*, 2021. arXiv: 2103.14030 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2103.14030>.
- [20] Z. Wang, Y. Zhu, H. Shi, Y. Zhang, and C. Yan, “A 3D multiscale view convolutional neural network with attention for mental disease diagnosis on MRI images,” en, *Math. Biosci. Eng.*, vol. 18, no. 5, pp. 6978–6994, Aug. 2021.
- [21] Z. S. Aaraji and H. H. Abbas, *Automatic classification of alzheimer’s disease using brain mri data and deep convolutional neural networks*, Mar. 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2204.00068>.
- [22] G. Andrade-Miranda, V. Jaouen, V. Bourbonne, F. Lucia, D. Visvikis, and P.-H. Conze, “Pure versus hybrid transformers for multi-modal brain tumor segmentation: A comparative study,” Oct. 2022, pp. 1336–1340. DOI: 10.1109/ICIP46576.2022.9897658.

- [23] M. A. Hassanien, V. K. Singh, D. Puig, and M. Abdel-Nasser, “Predicting breast tumor malignancy using deep convnext radiomics and quality-based score pooling in ultrasound sequences,” *Diagnostics*, vol. 12, no. 5, p. 1053, 2022. DOI: 10.3390/diagnostics12051053. [Online]. Available: <https://www.mdpi.com/2075-4418/12/5/1053>.
- [24] Z. K. Khadem-Reza and H. Zare, “Automatic detection of autism spectrum disorder (ASD) in children using structural magnetic resonance imaging with machine vision system,” en, *Middle East Curr. Psychiatr.*, vol. 29, no. 1, Dec. 2022.
- [25] Q.-H. Lin, Y.-W. Niu, J. Sui, W.-D. Zhao, C. Zhuo, and V. D. Calhoun, “Ssp-net: An interpretable 3d-cnn for classification of schizophrenia using phase maps of resting-state complex-valued fmri data,” *Medical Image Analysis*, vol. 79, p. 102430, 2022, ISSN: 1361-8415. DOI: <https://doi.org/10.1016/j.media.2022.102430>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1361841522000810>.
- [26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 11976–11986.
- [27] G. Tian, Z. Wang, C. Wang, *et al.*, “A deep ensemble learning-based automated detection of COVID-19 using lung CT images and vision transformer and ConvNeXt,” en, *Front. Microbiol.*, vol. 13, p. 1024104, Nov. 2022.
- [28] H. B. Baydargil, *Convnext-medical-imaging*, <https://github.com/HusnuBarisBaydargil/ConvNext-Medical-Imaging>, 2023.
- [29] Y. Bi, A. Abrol, Z. Fu, and V. D. Calhoun, “A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data,” *bioRxiv*, 2023. DOI: 10.1101/2023.07.14.549002. eprint: <https://www.biorxiv.org/content/early/2023/07/18/2023.07.14.549002.full.pdf>. [Online]. Available: <https://www.biorxiv.org/content/early/2023/07/18/2023.07.14.549002>.
- [30] A. Hamran, M. Vaeztourshizi, A. Esmaili, and M. Pedram, *Brain tumor detection using convolutional neural networks with skip connections*, Jul. 2023. DOI: 10.48550/arXiv.2307.07503.
- [31] M. Odusami, R. Maskeliūnas, and R. Damaševičius, “Pixel-level fusion approach with vision transformer for early detection of alzheimer’s disease,” en, *Electronics (Basel)*, vol. 12, no. 5, p. 1218, Mar. 2023.
- [32] A. Shanko, L. Abute, and T. Tamirat, “Attitudes towards schizophrenia and associated factors among community members in hossana town: A mixed method study,” *BMC Psychiatry*, vol. 23, no. 1, 2023, ISSN: 1471-244X. DOI: 10.1186/s12888-023-04555-9. [Online]. Available: <http://dx.doi.org/10.1186/s12888-023-04555-9>.
- [33] H. W. Wei, E. Cheng, and C. Peng, “Robustvision: Making cnn and vit good friends with pre-trained vision model,” Ph.D. dissertation, 2023.
- [34] W. Zhang, C. Yang, Z. Cao, *et al.*, “Detecting individuals with severe mental illness using artificial intelligence applied to magnetic resonance imaging,” en, *EBioMedicine*, vol. 90, no. 104541, p. 104541, Apr. 2023.

- [35] M. H. Alshayeji, “Alzheimer’s disease detection and stage identification from magnetic resonance brain images using vision transformer,” *Machine Learning: Science and Technology*, vol. 5, no. 3, p. 035 011, Jul. 2024. DOI: 10.1088/2632-2153/ad5fdc. [Online]. Available: <https://dx.doi.org/10.1088/2632-2153/ad5fdc>.
- [36] Y. Bi, A. Abrol, S. Jia, J. Sui, and V. D. Calhoun, “Gray matters: Vigan framework for identifying schizophrenia biomarkers linking structural mri and functional network connectivity,” *NeuroImage*, vol. 297, p. 120 674, 2024, ISSN: 1053-8119. DOI: <https://doi.org/10.1016/j.neuroimage.2024.120674>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1053811924001691>.
- [37] S. Khan, *A novel feature map enhancement technique integrating residual cnn and transformer for alzheimer diseases diagnosis*, Mar. 2024. DOI: 10.48550/arXiv.2405.12986.
- [38] J. Liu, S. Mao, and L. Pan, “Attention-based two-branch hybrid fusion network for medical image segmentation,” *Applied Sciences*, vol. 14, p. 4073, May 2024. DOI: 10.3390/app14104073.
- [39] Y. Mehmood and U. I. Bajwa, “Brain tumor grade classification using the ConvNext architecture,” in *Digit. Health*, vol. 10, p. 20 552 076 241 284 920, Jan. 2024.
- [40] F. J. Montalbo, L. Hernandez, L. Palad, R. Castillo, A. S. Alvin D.Eng, and A. L. De Ocampo, “Performance analysis of lightweight vision transformers and deep convolutional neural networks in detecting brain tumors in mri scans: An empirical approach,” Jan. 2024, pp. 17–25. DOI: 10.1145/3634875.3634878.
- [41] A. Rani Palakayala and P. Kuppusamy, “AttentionLUNet: A hybrid model for parkinson’s disease detection using MRI brain,” *IEEE Access*, vol. 12, pp. 91 752–91 769, 2024.
- [42] S. R. Siam, *Thesis_t2420320*, https://github.com/DragzterX/Thesis_T2420320, 2024.