Hierarchical Multi-Scale Deep Neural Network for Schizophrenia Detection in Neuroimaging

Carlos Dias Maia[1], Gabriel Barbosa da Fonseca[2], Luis Enrique Zárate Gálvez[1], Silvio Jamil Ferzoli Guimarães[2]

1- Applied Computational Intelligence Laboratory – LICAP,
Pontifical Catholic University of Minas Gerais, Brazil, 31980–110
2- Laboratory of Image and Multimedia Data Science (IMScience),
Pontifical Catholic University of Minas Gerais, Brazil, 31980–110

Abstract. Schizophrenia remains difficult to diagnose due to its reliance on subjective clinical assessment. This work proposes a pipeline for automated schizophrenia classification using functional MRI data from the UCLA CNP dataset. The method extracts multi-view slices from nine anatomical orientations using a hierarchical analysis and processes them with a Vision Transformer model (MultiSliceViT). Under stratified 5-fold cross-validation, the approach achieved 86.4% accuracy, outperforming models with fewer views. Interpretability analyses highlighted consistent attention to key regions, including the dorsolateral prefrontal cortex, hippocampus, and anterior cingulate. These results demonstrate the effectiveness of multi-view transformer architectures for identifying meaningful functional biomarkers.

1 Introduction

Schizophrenia is a complex psychiatric disorder that affects approximately 1% of the global population and is characterized by disturbances in thought, perception, emotion, and behavior [1]. Diagnosis remains predominantly clinical, based on symptomatological criteria and the subjective evaluation of mental health professionals, which poses a significant challenge due to the heterogeneity of the condition and the overlap of symptoms with other disorders [2].

Over the past decades, the integration of machine learning with neuroimaging has substantially advanced schizophrenia research. Early studies combining fMRI and artificial intelligence demonstrated that computational models can capture functional alterations in regions such as the prefrontal cortex, hippocampus, and the default mode network [3, 4], revealing neural markers not easily identifiable through clinical assessment alone. Subsequent developments in deep learning, particularly convolutional architectures, achieved strong performance in schizophrenia classification [5, 6].

More recent work has shifted toward transformer-based methods, which leverage self-attention [7] to model long-range spatial relationships. Vision Transformers (ViT) [8] and related architectures have shown promising results for both functional and structural neuroimaging, as highlighted in several surveys [9, 10]. In parallel, multi-view learning has emerged as an efficient strategy for capturing complementary anatomical information by extracting multiple 2D slices from 3D neuroimaging volumes [11]. Transformer-based multi-view fusion

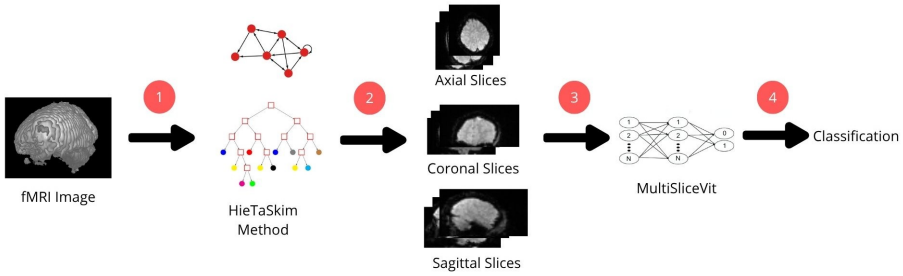

Fig. 1: Overview of the proposed processing pipeline.

methods such as TransFusion [12] further strengthen this paradigm by integrating information across axial, coronal, and sagittal planes, which is particularly valuable given the distributed nature of schizophrenia-related abnormalities.

Building on these developments, our work combines multi-view transformer modeling with an adaptive hierarchical slice-selection mechanism inspired by HieTaSkim [13]. The approach relies on the selective extraction of multiple two-dimensional views (slices) of specific brain regions from three-dimensional volumes, aiming to reduce the computational complexity inherent to full-volume fMRI processing without compromising diagnostic accuracy. We show in our experiments that with our best configuration we are able to achieve 86.4% accuracy using a stratified 5-fold cross-validation, using the UCLA Consortium for Neuropsychiatric Phenomics dataset [14].

2 Multi-scale ViT using hierarhies

This work proposes a multi-stage pipeline designed to efficiently extract informative anatomical representations from fMRI data and integrate them into a transformer-based multi-view classification model. The methodology consists of three main components: (i) preprocessing and slice-based reduction of 4D fMRI volumes, (ii) adaptive hierarchical selection of anatomically relevant slices using the HieTaSkim algorithm, and (iii) diagnostic classification through the proposed `MultiSliceViT` architecture (Figure 1). This design aims to balance computational cost with rich representation, combining efficiency and efficacy.

2.1 Preprocessing and Slice Extraction

The original 4D fMRI volumes consist of a temporal sequence of 3D acquisitions accompanied by demographic, physiological, and acquisition metadata. To reduce computational cost while preserving spatially relevant information, we adopt a slice-based representation in which two-dimensional images are extracted from the midpoint of the time series. This reduces each subject's data to a manageable set of structural views suitable for transformer-based processing.

Slice positions are determined through a hierarchical adaptive strategy derived from the HieTaSkim algorithm [13], originally developed for video summarization. In this adaptation, each fMRI volume is modeled as a hierarchical graph whose nodes correspond to candidate slice locations across sagittal, coronal, axial, and oblique orientations. Adaptive cuts identify the 18 most informative positions by balancing anatomical relevance and spatial diversity, enabling the automatic prioritization of critical regions such as the prefrontal cortex, hippocampus, and cingulate areas without requiring manual segmentation.

The selected multi-view representation allows the model to integrate complementary anatomical perspectives, capturing subtle morphological and functional variations that may be more prominent in different orientations. This approach provides an efficient compromise between representational richness and computational efficiency, facilitating large-scale analysis even in resource-limited settings.

2.2 MultiSliceViT Architecture

To classify schizophrenia from the extracted multi-view slices, we developed `MultiSliceViT`, a transformer-based architecture derived from the ViT-B/16 model [8]. The model is adapted to accept multiple grayscale slices as independent channels, replacing the traditional RGB input configuration. Let $\mathbf{X} \in \mathbb{R}^{B\times S\times C\times H\times W}$ denote the set of slices for each subject, where B represents the batch size, S the number of slices selected by HieTaSkim, $C = 1$ indicates grayscale images, and $H \times W$ the spatial dimensions of each slice; the tensor is reshaped to $\mathbf{X}_{reshape} \in \mathbb{R}^{B\times S\times H\times W}$ so that each slice is treated as an input channel.

The architectural pipeline comprises three stages. First, slices are divided into fixed-size patches and linearly projected into an embedding space, augmented with positional encodings:

$$\mathbf{Z}_0 = [\mathbf{x}_{class}; \mathbf{EP}_1; \ldots; \mathbf{EP}_N] + \mathbf{E}_{pos}. \tag{1}$$

Second, the patch embeddings are processed by a sequence of transformer blocks employing multi-head self-attention to capture long-range spatial relationships across anatomical locations. Finally, the model integrates information from sagittal, coronal, and axial views using a learnable linear fusion mechanism:

$$\mathbf{y}_{final} = \mathbf{W} \begin{bmatrix} \mathbf{y}_{axial} \\ \mathbf{y}_{sagittal} \\ \mathbf{y}_{coronal} \end{bmatrix} + \mathbf{b}, \tag{2}$$

where each $\mathbf{y}_{orientation} \in \mathbb{R}^d$ is the embedding derived from a specific anatomical orientation. This fusion strategy enables the model to exploit interplanar relationships and enhances its ability to detect distributed anomalies characteristic of schizophrenia.

Slices/Orientation	Accuracy	Sensitivity	Precision	F1-Score
2	74.9%	74.6%	75.2%	0.749
4	75.9%	76.2%	76.4%	0.761
8	78.8%	78.2%	78.0%	0.791
10	79.6%	79.8%	79.3%	0.790
12	80.8%	80.4%	80.7%	0.803
16	**82.6%**	**81.8%**	**83.4%**	**0.825**

Table 1: Progressive performance evaluation across different slice configurations

3 Results

3.1 Experimental Setup

The experiments were conducted on the UCLA Consortium for Neuropsychiatric Phenomics dataset [14], using stratified k-fold cross-validation ($k = 5$) to ensure statistical robustness and mitigate overfitting risks, considering the class imbalance (130 controls vs. 50 schizophrenia patients). Stratification ensured that each fold maintained similar proportions between classes, preserving the representativeness of the original dataset.

Training was performed with a batch size of 16, initial learning rate of 1×10^{-4} with exponential decay scheduler ($\gamma = 0.95$), Adam as optimizer, and early stopping based on validation loss with patience of 10 epochs. The cross-entropy loss function was weighted (weight 2.6 for schizophrenia, 1.0 for controls) to compensate for class imbalance. To avoid overfitting, spacial dropout was also used during training.

The performance of the proposed method was evaluated using five standard metrics commonly employed in machine learning and medical imaging: Accuracy, Sensitivity, Specificity, F1-Score, and AUC-ROC [15].

3.2 Progressive Evaluation of Multi-View Architecture

To assess the impact of anatomical coverage on diagnostic performance, we progressively varied the number of slices extracted per subject from 2 to 16. This experiment aimed to determine the minimum structural information required for the model to capture discriminative neuroanatomical patterns associated with schizophrenia. As shown in Figure 2 and Table 3.2, configurations with only 2 or 4 slices provided limited spatial context and achieved comparatively lower accuracies, with the 2-slice setup reaching 74.9%. These reduced configurations lacked sufficient anatomical variability for the model to reliably distinguish between patients and controls.

Performance improved consistently as more slices were incorporated. Intermediate configurations using 8, 10, or 12 slices benefited from richer anatomical coverage, allowing the model to capture more distributed information. The 12-slice configuration reached 80.8% accuracy, representing a substantial improvement over the 2-slice baseline. The highest performance was obtained with 16 slices, which achieved 82.6% accuracy and offered the most comprehensive

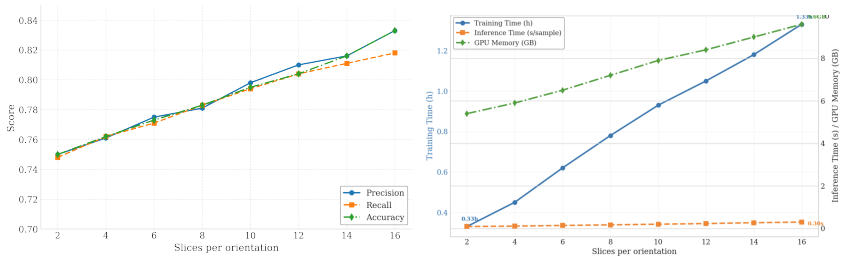

Fig. 2: Evolution of Performance Metrics, memory cost and time by Number of Slices

structural representation among all tested configurations.

Additionally, Figure 2 shows that, although memory usage and training time increase approximately linearly with the number of slices, the impact on inference time remains minimal. This indicates that the multi-view strategy is computationally feasible for real-world deployment.

3.3 Ablation and Interpretability Analysis

An ablation study was conducted to assess the relative contribution of each anatomical orientation within the multi-view model. The coronal plane exhibited the highest discriminative weight (41.34%), followed by the sagittal (32.26%) and axial (26.40%) orientations. This distribution is consistent with prior neuroimaging findings indicating that coronal views capture frontal and temporal regions frequently implicated in schizophrenia, thus providing more informative structural cues for classification.

To examine the neuroanatomical basis of the model's decisions, attention-based interpretability methods were applied. The attention maps indicated that the dorsolateral prefrontal cortex accounted for the highest share of model focus (34.2%), followed by hippocampal and other limbic structures (28.7%), the anterior cingulate cortex (19.3%), and subcortical regions including the thalamus and basal ganglia (17.8%). These regions correspond closely to established biomarkers reported in schizophrenia research, suggesting that the model prioritizes clinically relevant anatomical substrates.

4 Conclusion and future works

In this work, we introduced a hierarchical multi-scale pipeline for schizophrenia classification from fMRI data, combining an adaptive slice-selection strategy with a transformer-based multi-view architecture. By selecting multiple views through a hierarchical mechanism and processing them with the proposed MultiSliceViT model, the method achieved 86.4% accuracy under stratified 5-fold cross-validation, outperforming configurations with reduced anatomical coverage.

Interpretability analysis confirmed that the model concentrates on neurobiologically relevant regions implicated in schizophrenia, strengthening its clinical plausibility. Moreover, the small increase in inference time when increasing the number of slices highlights the practicality of the multi-view approach for real-world applications.

In future works, we plan to explore how different hierarchy structures and manipulation strategies influence classification performance. We also aim to integrate hierarchy construction directly into the learning loop, enabling task-adaptive slice selection.

References

[1] Dawn I. Velligan and Sanjai Rao. The epidemiology and global burden of schizophrenia. *J. Clin. Psychiatry*, 84(1):45094, 2023.

[2] Kathleen L. Benson and Irwin Feinberg. Chapter 131 - schizophrenia. pages 1501–1511, 2011.

[3] Jose A. Cortes-Briones, Nicolas I. Tapia-Rivas, Deepak Cyril D'Souza, and Pablo A. Estevez. Going deep into schizophrenia with artificial intelligence. *Schizophr. Res.*, 245:122–140, 2022. Computational Approaches to Understanding Psychosis.

[4] L.A. Whitten. *Functional magnetic resonance imaging (fMRI): An invaluable tool in translational neuroscience*. RTI Press, 2012.

[5] Juliet Polok Sarkar and András Hajdu. Comparative analysis of deep learning methods for schizophrenia classification from fmri scans. *CBMS*, pages 69–74, 2024.

[6] Takrouni Wiem and Douik Ali. Schizophrenia diagnosis from fmri data based on deep curvelet transform. In *SSD*, pages 35–40, 2021.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[9] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey, 2022.

[10] Vivens Mubonanyikuzo, Hongjie Yan, Temitope Emmanuel Komolafe, Liang Zhou, Tao Wu, and Nizhuan Wang. Detection of alzheimer disease in neuroimages using vision transformers: Systematic review and meta-analysis. *J Med Internet Res*, 27:e62647, Feb 2025.

[11] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning, 2013.

[12] Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, and Dimitris Metaxas. Transfusion: Multi-view divergent fusion for medical image segmentation with transformers, 2022.

[13] Leonardo Vilela Cardoso, July F.M. Werneck, Silvio Jamil F. Guimarães, and Zenilton K.G. Patrocínio. Unsupervised video skimming with adaptive hierarchical shot detection. In *SIBGRAPI*, pages 1–6, 2024.

[14] R Bilder, R Poldrack, T Cannon, E London, N Freimer, E Congdon, K Karlsgodt, and F Sabb. "ucla consortium for neuropsychiatric phenomics la5c study", 2018.

[15] David MW Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *J. Mach. Learn. Technol.*, 2(1):37–63, 2011.