Comparing MedSigLIP and Structured Connectivity Models for ADHD and Schizophrenia Classification

Eronides F. da Silva Neto[1], Beatriz L. Bonafini[2], Breno C. Bispo[1], Juliano B. Lima[1]

[1]Department of Electronics and Systems, Federal University of Pernambuco (UFPE)

[2]Centro de Estudos e Sistemas Avançados de Recife (CESAR School)

{eronides.silvaneto, breno.bispo, juliano.lima}@ufpe.br

blf@cesar.school

Abstract. *The diagnosis of ADHD and schizophrenia is evolving with the integration of functional magnetic resonance imaging and machine learning. This study compares structured connectivity models (graph and hypergraph) with MedSigLIP foundation-model embeddings for disorder classification. Subject-level representations were extracted from the ADHD-200 and COBRE datasets and evaluated using 5-fold cross-validation. Among four classifiers, SVM consistently achieved the best performance. For ADHD, SVM reached an mAUC of 64.89%, approaching a graph-based baseline. For schizophrenia, it achieved an mAUC of 62.48%. These results indicate a competitive alternative to handcrafted connectivity features, even without disorder-specific pretraining.*

1. Introduction

Attention-deficit/hyperactivity disorder (ADHD) is a common neurodevelopmental disorder affecting approximately 7.6% of children and adolescents and 4.4% of adults worldwide [Al-Wardat et al. 2024]. It is characterized by persistent inattention, hyperactivity, and impulsivity, often leading to functional impairments that extend into adulthood [Noah and Sedky 2025]. Schizophrenia (SZ), another major psychiatric disorder, typically emerges in late adolescence or early adulthood [Anwar et al. 2025]. Although it shares some cognitive features with ADHD, SZ is distinguished by positive symptoms (e.g., hallucinations and delusions) and negative symptoms, including emotional blunting and cognitive deficits [Anwar et al. 2025].

With the integration of computational technologies into medicine, data has become central to the development of artificial intelligence (AI) methods. Functional magnetic resonance imaging (fMRI) is a widely used noninvasive technique for assessing brain function, providing quantitative measures of neural activity and connectivity to support diagnosis and clinical decision-making [Du et al. 2024]. Advances in AI, driven by generative approaches [Achiam et al. 2023] and large language models (LLMs), have also led to domain-specific foundation models such as Google's MedGemma, designed to assist clinical workflows [Sellergren et al. 2025]. These developments motivate the evaluation of specialized vision models like MedSigLIP for high-dimensional neuroimaging to identify patterns associated with neurodevelopmental and psychiatric disorders.

Unlike general-purpose medical foundation models that rely on latent embeddings, neuroimaging-based diagnosis can use structured models that explicitly encode

brain organization [Xiao et al. 2020]. Graph and hypergraph models represent connectivity as edges and multiregional interactions as hyperedges, capturing pairwise and higher-order relationships among brain regions [Bassett and Sporns 2017]. These representations help identify regions and interaction patterns associated with model predictions in neurodevelopmental and psychiatric disorders.

Graph learning has also been extensively applied to ADHD and SZ. The Temporal-BCGCN model utilizes temporal graph convolutional networks and the DSF-BrainNet module to extract dynamic connectivity patterns and identify functional abnormalities in SZ [Zhu et al. 2024], while hypergraph-based methods capture higher-order interactions. For example, hypergraph foundation model (HGFM) leverages multidimensional pretraining and link prediction at individual and group levels [Han et al. 2025], and hypergraph attention networks model complex structural dependencies [Ji et al. 2022].

Despite advances in connectivity modeling and large-scale AI, few studies directly compare foundation-model embeddings with graph and hypergraph methods under identical datasets and tasks. To address this gap, we evaluate two paradigms for ADHD and schizophrenia classification: structured connectivity models and MedSigLIP-derived embeddings, providing a controlled comparison that quantifies differences in classification performance.

2. Materials and Methods

The overall pipeline (Fig. 1) illustrates the complete workflow adopted in this study, encompassing data preprocessing, embedding extraction, feature aggregation, and classification.

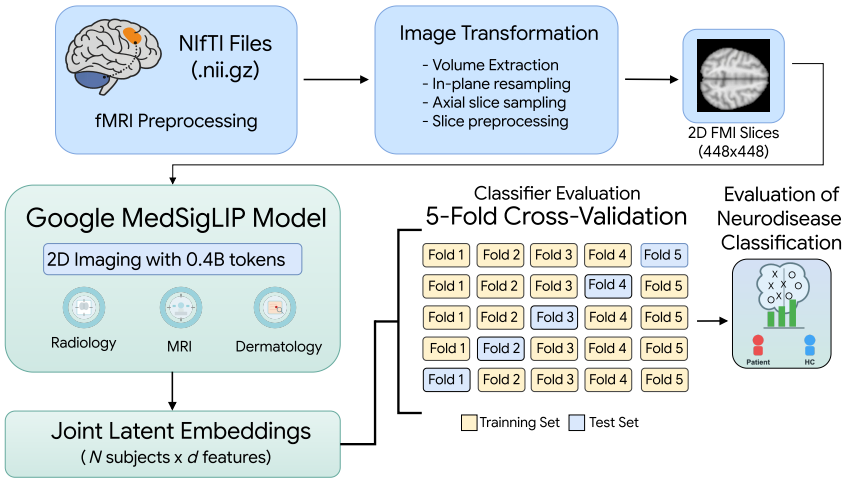

Figure 1. Overview of the methodological pipeline: fMRI preprocessing, 2D slice generation, embedding extraction with MedSigLIP, and evaluation via 5-fold cross-validation.

2.1. Datasets

The selected datasets for this work comprise real resting-state fMRI (rs-fMRI) data from the COBRE[1] and ADHD-200 [M. Milham, D. Fair, M. Mennes, S. Mostofsky 2012]

[1]The COBRE dataset is available at `https://figshare.com/articles/dataset/COBRE_preprocessed_with_NIAK_0_12_4/1160600`.

datasets. The COBRE dataset consists of preprocessed rs-fMRI scans from 72 individuals diagnosed with SZ and 74 healthy controls (HC), with participant ages ranging from 18 to 65 years. Data preprocessing was performed using the Neuroimaging Analysis Kit (NIAK).

For the ADHD-200 dataset, we used the New York University (NYU) cohort, which comprises preprocessed rs-fMRI data from children and adolescents diagnosed with ADHD and age-matched healthy controls, using the same preprocessing protocol adopted for the COBRE dataset. The subset includes 222 participants aged 7–17 years. Only participants diagnosed with the Combined (ADHD-C) and Hyperactive/Impulsive (ADHD-HI) subtypes were included, resulting in a final cohort of 220 subjects.

2.2. MedSigLIP

MedSigLIP is a medical-domain vision encoder developed by Google Research through large-scale vision–language pretraining, designed to learn transferable representations from medical images. Built on the SigLIP framework [Zhai et al. 2023], it produces aligned image–text embeddings that can be used in downstream tasks such as retrieval and classification. MedSigLIP uses dual vision and text encoders (400M parameters each) and is fine-tuned on over 33 million de-identified medical image–text pairs spanning multiple modalities (e.g., X-ray, CT, MRI, histopathology, ophthalmology, and dermatology).

In neurological contexts, the training data include datasets such as Slake-VQA, which contains radiology images with CT and MRI brain scans [Liu et al. 2021]. Images are processed at 448×448 resolution to generate high-dimensional embeddings for similarity and disease-related inference. In this work, MedSigLIP is employed as a pretrained feature extractor to generate latent embeddings, which are subsequently used to train a downstream supervised classifier. This embedding-based baseline is compared against graph and hypergraph-based methods that explicitly model relational structures.

2.3. Embeddings Calculation Procedure from fMRI Data

Because MedSigLIP requires fixed-resolution 2D inputs (448×448), the 4D rs-fMRI data were converted into an image-based representation. For each subject, 3D volumes were downsampled by a factor of 4 (i.e., one volume every four timepoints), preserving temporal coverage while reducing redundancy and retaining the original NIfTI affine for spatial consistency. Volumes were resampled in-plane to 448×448 using continuous interpolation, and a set of uniformly spaced axial slices was extracted, intensity-normalized to $[0, 1]$, rotated to radiological orientation, and converted to three-channel RGB images for model input.

Slice-level embeddings were computed with the pretrained MedSigLIP encoder and averaged to obtain a volume-level representation. Subject-level features were then generated by aggregating temporal embeddings using a learnable attention mechanism, yielding a fixed-length representation for classification. Mathematically, the slice-level embedding extraction process using the pretrained vision–language model is expressed as $\mathbf{z} = f_{\text{MedSigLIP}}(\mathbf{x}_{\text{brain image}})$, with $\mathbf{z} \in \mathbb{R}^d$, and $d = 1152$ fixed by the architecture of MedSigLIP vision encoder.

Let a subject have T sampled volumes, each represented by a volume-level embedding $\mathbf{v}_t \in \mathbb{R}^d$, for $t = 1, \ldots, T$. Given the set of embeddings associated with a subject,

the subject-level representation is computed as a weighted aggregation

$$\mathbf{z}_{\text{subj}} = \sum_{t=1}^{T} \alpha_t \mathbf{v}_t, \tag{1}$$

where α_t denotes the attention weight for the t-th volume and satisfies $\sum_{t=1}^{T} \alpha_t = 1$. Embedding extraction was performed in a cloud environment equipped with an NVIDIA Tesla T4 GPU (16 GB GDDR6), an Intel Xeon 2.20 GHz CPU, and 12 GB of system memory.

2.4. Experimental Evaluation

Using the same computational setup as in the embedding extraction stage, model evaluation follows a two-stage training and testing protocol with 5-fold cross-validation to ensure robustness. The dataset is partitioned into five folds; in each iteration, four folds are used for training and one for testing. Within each training split, LASSO is employed for feature selection to identify the most informative predictors and reduce dimensionality prior to classification. Hyperparameters are then optimized using Optuna. The model is subsequently retrained with the selected features and optimized configuration on the full training data and evaluated on the held-out fold to assess generalization performance.

Four supervised learning architectures: k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP) are evaluated. Performance is assessed using accuracy, precision, recall, F1-score and mean area under the ROC curve (mAUC). While accuracy and F1-score provide balanced measures from the confusion matrix to account for potential class imbalance, the mAUC offers a threshold-independent assessment of discriminative ability.

3. Experimental Results and Discussion

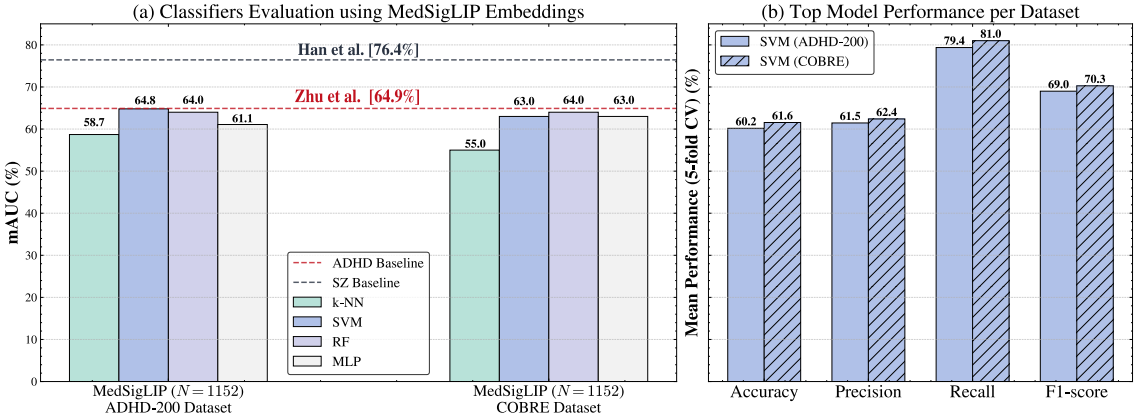

Figure 2. Performance of classifiers using MedSigLIP embeddings on ADHD-200 and COBRE. (a) Mean AUC. (b) Best model performance across metrics (5-fold cross-validation).

Figure 2 provides a summary of our experimental results. SVM and RF achieved the highest mAUC values under the 5-fold cross-validation protocol. The analysis prioritizes mAUC due to its threshold-independent assessment of discriminative performance

across all operating points, enabling a more robust comparison between classifiers, particularly in the presence of class imbalance or varying decision thresholds where accuracy alone may be insufficient.

The SVM achieved an accuracy of $(60.18 \pm 4)\%$ and an mAUC of $(64.89 \pm 7)\%$, indicating moderate discriminative performance. The recall was $(79.37 \pm 10)\%$, evidencing higher sensitivity, whereas precision was lower at $(61.5 \pm 2.7)\%$, resulting in an F1-score of $(69 \pm 4.6)\%$. These results suggest that SVM favors sensitivity, which may be advantageous in screening scenarios where minimizing false negatives is critical.

Similarly, SVM achieved the best performance on the COBRE dataset. Under the cross-validation protocol, the SVM model obtained an mAUC of $(62.48 \pm 12.03)\%$ and an accuracy of $(61.56 \pm 12.39)\%$, indicating moderate discriminative capability. In contrast to ADHD-200, where higher recall was observed, COBRE presented a more balanced trade-off between precision $(62.37 \pm 11.72)\%$ and recall $(59.62 \pm 11.44)\%$, yielding an F1-score of $(60.61 \pm 11.09)\%$. The larger standard deviations further indicate greater variability and reduced stability across folds compared to ADHD-200.

Compared to graph-based methods, MedSigLIP embeddings with SVM achieve competitive classification performance across both datasets, reaching an mAUC of 64.89% on ADHD-200 against the Zhu et al. baseline of 64.9%, and 62.48% on COBRE against the Han et al. benchmark of 76.4%. While structured connectivity models retain an advantage on schizophrenia classification, particularly in stability, given the higher variance observed on COBRE, MedSigLIP narrows this gap considerably for ADHD without relying on domain-specific graph construction. Notably, the SVM-MedSigLIP combination yields higher recall than precision on ADHD-200 (79.37% vs. 61.5%), suggesting a sensitivity profile more favorable for clinical screening than typical graph-based classifiers. These results indicate that foundation-model embeddings represent a viable, structurally agnostic alternative to handcrafted connectivity features, despite having no exposure to SZ or ADHD data during pretraining.

4. Conclusion

This study compared structured connectivity models and MedSigLIP embeddings for ADHD and schizophrenia classification. Results show that foundation-model embeddings achieve competitive performance, particularly with SVM for ADHD. These findings suggest that medical foundation-model representations are a viable alternative to handcrafted connectivity methods for neuroimaging-based psychiatric classification. Given the limited exposure of MedSigLIP to fMRI data during pretraining, disease-specific fine-tuning represents a promising direction for improving domain adaptation and model performance.

Acknowledgments

The authors wish to thank CNPq under grants 40151/2022-2, 442238/2023-1, 312935/2023-4 and 405903/2023-5, as well as CESAR School for institutional support.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Al-Wardat, M., Etoom, M., Almhdawi, K. A., Hawamdeh, Z., and Khader, Y. (2024). Prevalence of attention-deficit hyperactivity disorder in children, adolescents and adults in the middle east and north africa region: a systematic review and meta-analysis. *BMJ open*, 14(1):e078849.

Anwar, A., Mustafa, A. M., Abdou, K., A. Rabie, M., El-Shiekh, R. A., and El-Dessouki, A. M. (2025). A comprehensive review on schizophrenia: epidemiology, pathogenesis, diagnosis, conventional treatments, and proposed natural compounds used for management. *Naunyn-Schmiedeberg's Archives of Pharmacology*, pages 1–25.

Bassett, D. S. and Sporns, O. (2017). Network neuroscience. *Nature neuroscience*, 20(3):353–364.

Du, Y., Fang, S., He, X., and Calhoun, V. D. (2024). A survey of brain functional network extraction methods using fmri data. *Trends in Neurosciences*, 47(8):608–621.

Han, X., Xue, R., Feng, J., Feng, Y., Du, S., Shi, J., and Gao, Y. (2025). Hypergraph foundation model for brain disease diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17702–17716.

Ji, J., Ren, Y., and Lei, M. (2022). Fc–hat: Hypergraph attention network for functional brain network classification. *Information Sciences*, 608:1301–1316.

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., and Wu, X.-M. (2021). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

M. Milham, D. Fair, M. Mennes, S. Mostofsky (2012). The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62.

Noah, A. A. and Sedky, H. E. (2025). New frontiers in pharmacological treatment of attention-deficit hyperactivity disorder. *Naunyn-Schmiedeberg's Archives of Pharmacology*, pages 1–11.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al. (2025). Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Xiao, L., Wang, J., Kassani, P. H., Zhang, Y., Bai, Y., Stephen, J. M., Wilson, T. W., Calhoun, V. D., and Wang, Y.-P. (2020). Multi-hypergraph learning-based brain functional connectivity analysis in fmri data. *IEEE Transactions on Medical Imaging*, 39(5):1746–1758.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training.

Zhu, C., Tan, Y., Yang, S., Miao, J., Zhu, J., Huang, H., Yao, D., and Luo, C. (2024). Temporal dynamic synchronous functional brain network for schizophrenia classification and lateralization analysis. *IEEE Transactions on Medical Imaging*, 43(12):4307–4318.