BrainSCK: Brain Structure and Cognition Alignment via Knowledge Injection and Reactivation for Diagnosing Brain Disorders

Lilong Wang[1], Mianxin Liu[1], Shaoting Zhang[1], Xiaosong Wang[1]

[1]Shanghai AI Laboratory

Abstract. Emerging evidence from advanced neuroimaging study suggests common neurological bases across different brain disorders (BD) throughout the human lifespan. Researchers thus aim to create a general neuroimaging-based diagnosis model for population-scale screening for multiple BDs. Existing models predominantly use the transfer learning paradigm for BD tasks based on either out-of-domain models pre-trained with large-scale but less-related data and tasks or in-domain models pre-trained on healthy population brain data with auxiliary tasks such as age prediction. The former approach has few recognition of inter-individual variations and BD-related features in the population-scale brain data, while the latter relies on weak implicit association between the proxy and BD tasks. In this work, we propose a two-stage vision-language model adaptation strategy to incorporate novel knowledge into the out-of-domain well pre-trained model (e.g., BLIP) by aligning basic cognition and brain structural features for accurate diagnosis of multiple BDs. First, using life-span Human Connectome Project data, we textualize the demographics and psychometrics records and construct knowledge-injecting textual prompts (with important cognitive science contexts). The model is expected to learn the alignment between brain structure from images and cognitive knowledge from texts. Then, we customize knowledge-reactivating instructions and further tune the model to accommodate the cognitive symptoms in each BD diagnosis task. Experimental results show that our framework outperforms other state-of-the-art methods on three BD diagnosis tasks of different age groups. It demonstrates a promising and feasible learning paradigm for adapting large foundation models to the cognitive neuroscience and neurology fields. Code is available at https://github.com/openmedlab/BrainSCK.

Keywords: Brain disorder · Vision-language model · Cognition · Knowledge-injecting prompt · Knowledge-reactivating instruction

Lilong Wang and Mianxin Liu contribute equally.
Corresponds to wangxiaosong@pjlab.org.cn (Xiaosong Wang).

1 Introduction

The mental health of children and elders is frequently affected by a wide spectrum of brain disorders (BDs), such as attention deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), and Alzheimer's disease (AD), whose prevention, diagnosis, and treatment remain challenging. Recently, there has been accumulating evidence suggesting that different BDs in early and late life could share certain common cognitive symptoms and brain structural and functional bases[20,19,17]. Researchers thus attempt to locate the transdiagnostic biomarkers and build a generalized foundation model for facilitating the diagnosis of multiple BDs based on neuroimaging data, whose success would reform current diagnosis and treatment frameworks of multiple BDs.

Existing models are mainly utilizing two streams of training framework. Some models are pre-trained on natural images or medical images for other organs with general supervision tasks and tuned directly for classifying healthy and BD populations [22,3,16]. This routine aims to utilize the potential of models pre-trained on rich data resources but could ignore neural processes associated with the BD-related alternations, such as aging effects, gender effects, and individual differences in cognitive ability. These alternations could interfere with the detection of general health-to-BD deviation and confuse the model. Alternatively, more models are first established for the normal neural processes in brain data from a healthy population with pre-training tasks, such as age prediction, gender prediction, and cognitive ability prediction[1,9,23,13]. Then, the pre-trained models are transferred to identify BDs. This stream of methods theoretically relies on the association between the pre-training task and the BD, e.g., the relationship between aging and AD. However, they remain limited in systematically modeling the various contexts and knowledge of cognitive functions when using simple numbers to encode the task goal. In addition, as only image representations are shared, the lack of explicit activation of the associated features of BDs from the pre-trained model could vastly suppress the usefulness of a pre-trained model and lead to less effectiveness in diagnosing BDs.

The standardized toolkit, NIH toolbox [26], has been popular in the field of cognitive neuroscience. It systematically measures basic cognitive abilities that span the general ability, including executive function, episodic memory, language, processing speed, etc. These scales are closely related to the hallmark cognitive alternations under BDs, such as the executive function deficits in ASD and ADHD [10,18]. It will be of paramount value to translate these measurements in the contexts of cognitive functions and associate them with the brain structure in imaging data. Inspired by the recent advance of vision-language models,e.g., BLIPs [15,14], we align the basic cognition with brain structural features in images to complement the model pre-trained with nature image-text pairs. In such a training setting, it can facilitate the precise reactivation of pre-learned knowledge in downstream BD analysis using prompt-based methods, e.g.,[27].

In this work, we propose a novel pre-training framework to align basic cognition and brain structural features and intend to build a general pre-trained model for diagnosing multiple BDs throughout the human lifespan. By employ-

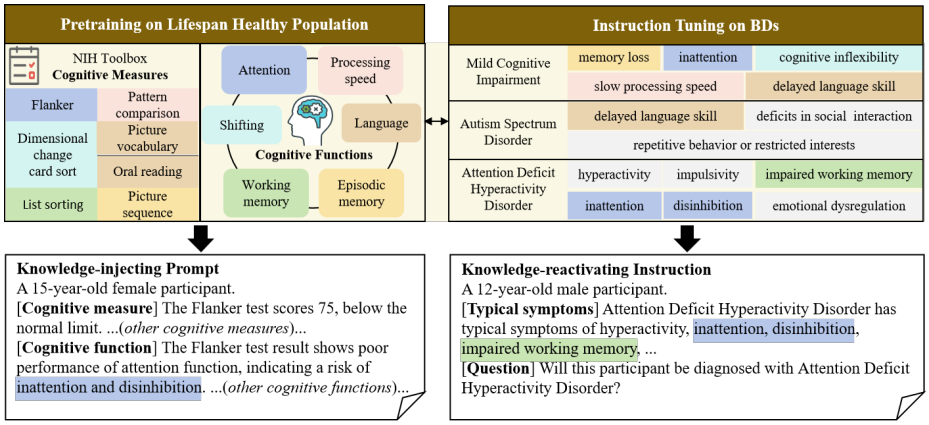

Fig. 1. Demonstrations of the proposed knowledge-injecting prompt and knowledge-reactivating instruction for model pretraining and downstream tuning, respectively, inspired by the conceptual dependencies (color-coded) among cognitive function, measures, and symptoms.

ing the population-scale human connectome project (HCP) data [2,21,24] with a variety of lifespans, we adapt the BLIP-2 model to learn further the alignment between brain structural T1 images and the knowledge-injecting prompts (textual tokens, shown in Fig. 1) composed with image-associated subjects' age, gender, and the NIH toolbox psychometrics. Additionally, we design knowledge-reactivating instruction for the instruction tuning stage, letting the model recall the pre-learned knowledge regarding the target BD diagnosis tasks. Using the proposed framework, our model demonstrates higher performance in diagnosing ASD, ADHD, and early diagnosis of AD than other state-of-the-art (SOTA) methods, even when our method only uses 30% of the training data.

2 Materials and Methods

The proposed framework based on Brain Structure, Cognition, and Knowledge, named BrainSCK, involves two stages of pre-training and instruction tuning, respectively. In the first stage, BrainSCK aligns the brain structure to demographic and cognitive information, formatted as knowledge-injecting prompts, using the healthy subjects from lifespan HCP data. The second stage involves teaching the model to conduct BD diagnosis as a visual question-answering (VQA) task with the help of knowledge-reactivating instructions. The BrainSCK model largely leverages the BLIP-2 [14] architecture to process vision-language data.

2.1 BrainSCK Model Architecture

As depicted in Fig. 2 (c), the model consists of a 3D projector, a pre-trained image encoder, and a querying transformer. Initially, the volumetric input im-

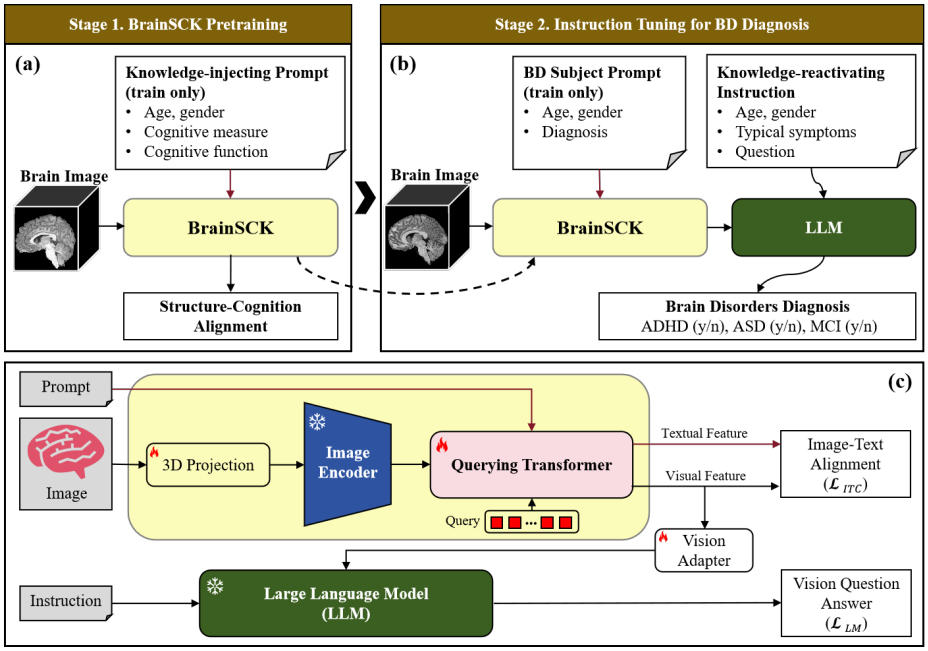

Fig. 2. Overview of the BrainSCK framework: (a) and (b) the workflows of pre-training and instruction tuning; (c) Details about the BrainSCK model.

age I is segmented into a series of 3D blocks. Similar to the pipeline in ViT [7], these blocks are transformed into embedding sequences using a learnable linear projection and added with position embeddings. These embedding sequences pass through the frozen image encoder f_v to form the image embeddings. In the querying transformer (Q-Former, learnable) f_q, the image embeddings interact with k learnable query vectors Q through cross-attention, resulting in k visual features V. Concurrently, the prompt undergoes a transformation into textual embeddings T using the text encoder of Q-Former. The resulting visual features V are then aligned with the textual embeddings T using the image-text contrastive loss $\mathcal{L}_{ITC}$.

For the instruction tuning stage, a fixed large language model (LLM, we use BioMedLM in this work) f_{lm} is employed after the BrainSCK model to receive both the adapted visual encodings $V^{'}$ from the Q-Former and the task instruction S as input to generate the answer A guided by the language modeling loss $\mathcal{L}_{LM}$.

2.2 Pretraining

Data preparation We select the cross-section data in HCP-D (development), HCP-YA (young adult), and HCP-A (aging) from the life-span HCP dataset to pre-train the model. The dataset thus covers 2,210 subjects aged from 6 to 86 years old and allows learning about normal development and aging in the large

span of human life. The details of the data characteristics are listed in Supplementary Table 1. The T1-weighted images provide structural features of the individual brain, which is preprocessed with an HCP minimal preprocessing pipeline [8] and is available from the HCP website. The demographic information such as age and gender is also included. Simultaneously, we extract seven age-adjusted standard scores of NIH toolbox psychometric testing from the corresponding subjects. Each of the psychometrics aims to measure a specific cognitive ability. The details and correspondence can be found in Fig. 1 as color codings.

Knowledge-injecting prompt The knowledge-injecting prompt is then designed for demographics, especially psychometrics. For demographics, we directly fill in the values into a text template. For psychometrics, a score falling below 85 indicates the relatively poor performance of the related cognitive function by design, and vice versa [26]. Therefore, besides the value-filling template, as "...scores..., below..." or "...scores..., above...", we additionally explain the values as either "...shows poor performance...indicating..." or "...shows normal performance..." (see Fig. 1 and Supplementary Table 2). For each psychometric, the wording to explain the poor performance is separately and carefully designed based on related cognitive neuroscience literature and the potential links to BD symptoms.

Structure-cognition alignment All psychometric modalities, rather than related ones, from the entire HCP dataset, are utilized for pretraining. The knowledge-injecting prompts, featuring a comprehensive description of cognitive abilities alongside brain images, are input into the BrainSCK model for alignment training. The training process concludes when the loss $\mathcal{L}_{ITC}$ converges.

2.3 Instruction Tuning

Knowledge-reactivating instruction To facilitate BD diagnosis, we craft knowledge-reactivating instructions comprising demographics (age, gender), typical BD symptoms, and a binary question. To activate and enhance the connection between pre-training and downstream tasks, we incorporate key cognitive symptoms from the BD symptom list, using identical wording as that used to describe healthy subjects with relatively poor cognitive abilities in the knowledge-injecting prompt. As shown in Fig. 1 and Supplementary Table 3, the typical symptoms outlined in the instruction for ADHD involve "inattention", "disinhibition" and "impaired working memory", which are closely linked to cognitive functions such as attention and working memory.

VQA for BD diagnosis During instruction tuning, the BrainSCK module is initialized with the pre-trained parameters. Subsequently, image-text alignment is conducted between the brain image and corresponding BD subject prompt, including age, gender, and diagnosis (Fig. 2 (b)). Following this, we employ

knowledge-reactivating instructions to guide the LLM model in diagnosing specific BD alongside the aligned visual features. In this stage, we use both $\mathcal{L}_{ITC}$ and $\mathcal{L}_{LM}$. During inference, the model can offer the diagnosis answer with a paired image-instruction input.

3 Experimental details

3.1 BD datasets

The ADHD-200 [5] and Autism Brain Imaging Data Exchange-I (ABIDE-I) [6] datasets for ADHD and ASD studies are used for early-life BD identifications. From ADHD-200, we include T1 data, age, and gender from 766 subjects (279 ADHD). And for ABIDE-I, we select 327 autism and 492 typical development subjects with the same data modality. AD is chosen as the representative BD in the late life of humans. An early AD diagnosis is essential but more difficult. Therefore, from Alzheimer's disease neuroimaging initiative (ADNI) [11], we include the latest scan of image and demographic data from 683 subjects with 317 mild cognitive impairment (MCI), which is regarded as an early status of AD. More information about the three datasets can be found in Supplementary Table 1. Each of these datasets is randomly divided into training, validation, and testing subsets using a 7:1:2 ratio.

3.2 Implementation

In the pretraining stage, we maintain the same frozen image encoder (ViT-g/14 [7]) and Q-Former (initialized with BERT-base [12]) as in BLIP-2. During the instruction tuning stage, the Q-Former inherits from our pre-trained version. The LLM of FLAN-T5 [4] used in BLIP-2 is replaced by a medical field LLM of BioMedLM [25]. During training, we set the batch size to 4 and utilized the AdamW optimizer with a learning rate of 2×10^{-5}. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

3.3 Evaluation metrics

As our model only generates diagnosis as texts rather than probability, we used three metrics to evaluate the model performance in BD diagnosis, including accuracy, F1-score, and Cohen's κ. F1-score is a metric that combines precision and recall, assessing the balance between false positives and false negatives. κ is a statistical measure that assesses the agreement between predicted and actual classification beyond what would be expected by chance.

4 Results

4.1 Comparison study

In Table 1, we compare our proposed method with different SOTA pre-trained methods on three BD diagnosis tasks. Our method demonstrates improvements

Table 1. Comparisons between four existing methods with the proposed BrainSCK for BD diagnosis on the testing subsets of ADHD-200, ABIDE-I, and ADNI datasets.

Method	ADHD-200			ABIDE-I			ADNI		
	Acc	F1	κ	Acc	F1	κ	Acc	F1	κ
Med3D[3]	0.669	0.541	0.289	0.642	0.337	0.136	0.642	0.462	0.211
UniFormer[16]	0.695	0.544	0.316	**0.661**	0.378	0.183	0.640	**0.608**	0.288
DeepBrainNet[1]	0.675	0.537	0.292	0.642	0.416	0.174	0.642	0.566	0.262
BLIP-2[14]	0.662	0.519	0.263	0.594	**0.472**	0.143	0.620	0.500	0.196
BrainSCK	**0.701**	**0.549**	**0.327**	0.636	0.454	**0.187**	**0.650**	0.603	**0.312**

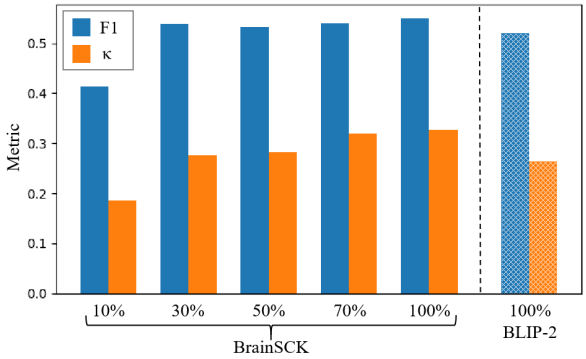

Fig. 3. Performance of BrainSCK using different ratios of training samples for diagnosing ADHD.

compared to BLIP-2[14] (assisted with BioMedLM for VQA) without domain data and knowledge pretraining. Moreover, our method outperforms other methods, including UniFormer [16] and Med3D [3] from general vision and medical fields which are based solely on vision information (T1), as well as deep brain network transferred from brain age prediction[1], across most of the evaluation metrics. Note the absolute levels of metrics are not high. We attribute this to the multi-site effects in ADHD and ASD data, the difficulty of MCI diagnosis, and the limited pre-training data size.

To demonstrate the potential of our method under small-sample conditions, we further evaluate the performance of BrainSCK using different ratios of training samples (from 10% to 70%) for ADHD diagnosis. Only F1 and κ values are shown due to the imbalanced classifications. In Fig. 3, we can observe that utilizing 30% of the samples leads to notable improvement of approximately 0.12 and 0.09 in F1 and κ values, respectively, compared to using only 10% of the samples. As the number of training samples continues to increase, the model's performance gradually approaches the one achieved with the entire training set.

Remarkably, even when utilizing only 30% of the training samples, our method surpasses the performance of BLIP-2 using all available training samples.

4.2 Ablation study

In Table 2, we change the configuration in pretraining and instruction tuning and re-perform the ADHD identification using 30% training data to assess the effectiveness of our design. Firstly, in the pre-training stage, we remove either HCP-D or HCP-A to highlight the importance of using life-span training data. It can be observed that both data removal results in a decrease in F1-score and κ, indicating that comprehensive learning of life-span brain changes, rather than solely focusing on early-age development, contributes to the identification of ADHD. Secondly, we get rid of the cognitive function descriptions from the knowledge-injecting prompt. As a consequence, both F1 and κ are remarkably undermined. This supports the cognitive function descriptions as one of the core designs in the knowledge-injecting prompt. Additionally, the instruction content is modified. Removing either demographic information or disease symptom description from the instruction undermines the diagnosis performance reflected in all metrics, highlighting the significance of the entire design of the knowledge-reactivating instruction. Lastly, we observe that BioMedLM is a better choice for BrainSCK than FLAN-T5, emphasizing the importance of medical domain training.

Table 2. Ablation experimental results on the ADHD-200 dataset under different settings.

Setting	Option	F1	κ
HCP Dataset	w/o HCP-D	0.454	0.202
	w/o HCP-A	0.521	0.237
Prompt	w/o cognitive function description	0.508	0.209
Instruction	w/o age, gender	0.486	0.205
	w/o typical symptoms	0.477	0.197
Language Model	FLAN-T5	0.306	0.047
Proposed	BrainSCK	**0.539**	**0.276**

5 Conclusion

We propose a novel two-stage vision-language training framework for multiple BD diagnosis. Our design of knowledge-injecting prompts contributes to the alignment between brain structural features and cognitive functions. The proposed knowledge-reactivating instruction enhances the connection between the

learned knowledge about cognition and the BD cognitive symptoms, facilitating the diagnosis. Our framework demonstrates superior diagnostic accuracy for ASD, ADHD, and early-stage AD compared to state-of-the-art methods, even under limited fine-tuning data. In general, our work bridges foundation model, cognitive neuroscience, and clinical studies and offers a promising tool for the diagnosis of a broad spectrum of BDs throughout the human lifespan. The proposed method is currently preliminary, and the performance is limited by training sample size and sub-optimal prompt design. We believe it has space for further improvements and more comprehensive evaluations.

Acknowledgments

This work is funded by the National Key R&D Program of China (2022ZD0160700) and Shanghai AI Laboratory.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bashyam, V.M., Erus, G., Doshi, J., Habes, M., Nasrallah, I.M., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., et al.: Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. Brain **143**(7), 2312–2324 (2020)
2. Bookheimer, S.Y., Salat, D.H., Terpstra, M., Ances, B.M., Barch, D.M., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Diaz-Santos, M., Elam, J.S., et al.: The lifespan human connectome project in aging: an overview. Neuroimage **185**, 335–348 (2019)
3. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
4. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
5. consortium, A..: The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in systems neuroscience **6**, 62 (2012)
6. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry **19**(6), 659–667 (2014)
7. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)

8. Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., et al.: The minimal preprocessing pipelines for the human connectome project. Neuroimage **80**, 105–124 (2013)
9. He, T., An, L., Chen, P., Chen, J., Feng, J., Bzdok, D., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.: Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. Nature Neuroscience **25**(6), 795–804 (2022)
10. Hill, E.L.: Executive dysfunction in autism. Trends in cognitive sciences **8**(1), 26–32 (2004)
11. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
12. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2 (2019)
13. Kim, P., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., Yoo, S., Cha, J., Moon, T.: Swift: Swin 4d fmri transformer. Advances in Neural Information Processing Systems **36** (2024)
14. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2023)
16. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
17. Liu, M., Zhang, J., Wang, Y., Zhou, Y., Xie, F., Guo, Q., Shi, F., Zhang, H., Wang, Q., Shen, D.: A common spectrum underlying brain disorders across lifespan revealed by deep learning on brain networks. iScience **26**(11), 108244 (2023)
18. Martel, M., Nikolas, M., Nigg, J.T.: Executive function in adolescents with adhd. Journal of the American Academy of Child & Adolescent Psychiatry **46**(11), 1437–1444 (2007)
19. Menon, V.: Large-scale brain networks and psychopathology: a unifying triple network model. Trends in cognitive sciences **15**(10), 483–506 (2011)
20. Opel, N., Goltermann, J., Hermesdorf, M., Berger, K., Baune, B.T., Dannlowski, U.: Cross-disorder analysis of brain structural abnormalities in six major psychiatric disorders: a secondary analysis of mega-and meta-analytical findings from the enigma consortium. Biological Psychiatry **88**(9), 678–686 (2020)
21. Somerville, L.H., Bookheimer, S.Y., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Dapretto, M., Elam, J.S., Gaffrey, M.S., Harms, M.P., Hodge, C., et al.: The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. Neuroimage **183**, 456–468 (2018)
22. Talo, M., Yildirim, O., Baloglu, U.B., Aydin, G., Acharya, U.R.: Convolutional neural networks for multi-class brain disease detection using mri images. Computerized Medical Imaging and Graphics **78**, 101673 (2019)

23. Tanveer, M., Ganaie, M., Beheshti, I., Goel, T., Ahmad, N., Lai, K.T., Huang, K., Zhang, Y.D., Del Ser, J., Lin, C.T.: Deep learning for brain age estimation: A systematic review. Information Fusion (2023)
24. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.M.H., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
25. Venigalla, A., Frankle, J., Carbin, M.: Biomedlm: a domain-specific large language model for biomedical text. MosaicML. Accessed: Dec **23**(3), 2 (2022)
26. Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P.D., Bauer, P.J., Carlozzi, N.E., Slotkin, J., Blitz, D., Wallner-Allen, K., et al.: Cognition assessment using the nih toolbox. Neurology **80**(11 Supplement 3), S54–S64 (2013)
27. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)