Enhancing OCD classification with Transformer-based deep learning on resting-state fMRI: insights from the ENIGMA-OCD cohort and UK Biobank pretraining

Maria Pak[1], Youngchan Ryu[2], Sangyoon Bae[3], Willem B. Bruin[5,6], Guido A. van Wingen[5,6], Odile A. van den Heuvel[7,8,9], Jiook Cha[1,3,4], ENIGMA-OCD Working Group

[1] Department of Psychology, Seoul National University, Seoul, Republic of Korea

[2] Department of Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea

[3] Graduate School of Artificial Intelligence, Seoul National University, Seoul, Republic of Korea

[4] Department of Brain and Cognitive Sciences, Seoul National University, Seoul, Republic of Korea

[5] Amsterdam UMC location University of Amsterdam, Department of Psychiatry, Meibergdreef 9, Amsterdam, The Netherlands

[6] Amsterdam Neuroscience, Amsterdam, The Netherlands

[7] Amsterdam UMC, location Vrije Universiteit Amsterdam, Department of Psychiatry, De Boelelaan 1117, Amsterdam, The Netherlands

[8] Amsterdam UMC, location Vrije Universiteit Amsterdam, Department of Anatomy and Neurosciences, De Boelelaan 1117, Amsterdam, The Netherlands

[9] Amsterdam Neuroscience, Compulsivity, Impulsivity & Attention program, Amsterdam, The Netherlands

Abstract

Obsessive-compulsive disorder (OCD) remains challenging to classify due to its heterogeneous clinical presentation and the limitations of static brain connectivity metrics. To address these hurdles, we applied a Transformer-based deep learning model to a record-sized dataset of resting-state fMRI from 2,094 individuals in the ENIGMA-OCD consortium. By pretraining on an extensive UK Biobank dataset and using dynamic connectivity measures across multiple frequency bands, our approach achieved higher predictive performance for OCD than conventional methods. We further conducted uncertainty quantification, revealing a marked reduction in calibration error for the pretrained model. Finally, self-attention-based interpretation pinpointed reduced connectivity within sensorimotor networks in patients with OCD, consistent with prior literature. These findings underscore the value of large-scale pretraining and dynamic rs-fMRI data in enhancing model generalizability, highlighting a promising avenue for more robust OCD classification and, by extension, clinical decision-making.

Keywords: OCD classification, deep learning, fMRI, dynamic functional connectivity.

Introduction

Obsessive-compulsive disorder (OCD) is a psychiatric disorder that remains challenging to diagnose and treat, with its complex neurobiological underpinnings not fully understood. Prior OCD studies have been limited by small sample sizes and homogeneous datasets, hindering the generalizability of their findings (Bruin, Denys & van Wingen, 2019). Moreover, the predominant focus on **static functional connectivity** from resting-state fMRI (rs-fMRI) neglects time-varying changes that capture the brain's intrinsic dynamics. To overcome these hurdles, we propose a **Transformer-based multi-band approach** that explicitly models dynamic connectivity. By **pretraining** on the extensive **UK Biobank** resource (40,708 participants), we aim to imbue our OCD-specific model with broader representational capacity. We further disentangle frequency-specific contributions in brain networks using **variational mode decomposition**, hypothesizing that this multi-band strategy will yield more robust OCD classification.

Methods

Dataset. We used rs-fMRI from the ENIGMA-OCD cohort, which includes 29 international sites comprising 2,094 participants (1,040 OCD patients and 1,054 healthy controls). For pretraining, we utilized UK Biobank data, which includes 40,708 participants, enabling our model to learn from a large-scale, generalizable sample.
Model. We applied a Transformer-based Multi-Band Brain Net (MBBN) model to the fMRI time-series data, dividing it into four frequency bands using variational mode decomposition (Fig. 1; Bae et al., 2025; Dragomiretskiy & Zosso, 2014). Frequency band cutoffs were calculated for each subject. We pretrained the MBBN using masked signal modeling on the UK Biobank data to enhance predictive accuracy. Leave-one-site-out cross-validation (LOSO-CV) was employed to assess generalizability across sites. To quantify model uncertainty, we used Monte Carlo (MC) dropout, which estimates how well the model's confidence aligns with the empirical accuracies (Srivastava et al., 2014). Finally, we leveraged Grad-CAM-based attribution (Selvaraju et al., 2017) and self-attention matrices to identify significant functional connections associated with OCD classification.

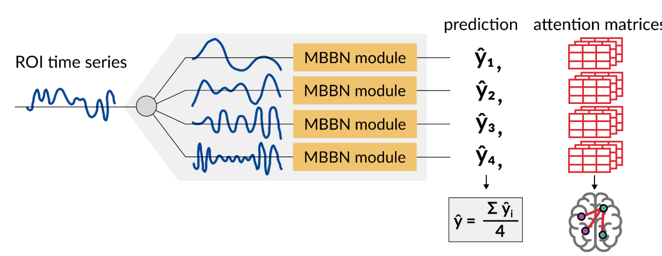

Figure 1: Model architecture of the Multi-Band Brain Net (MBBN).

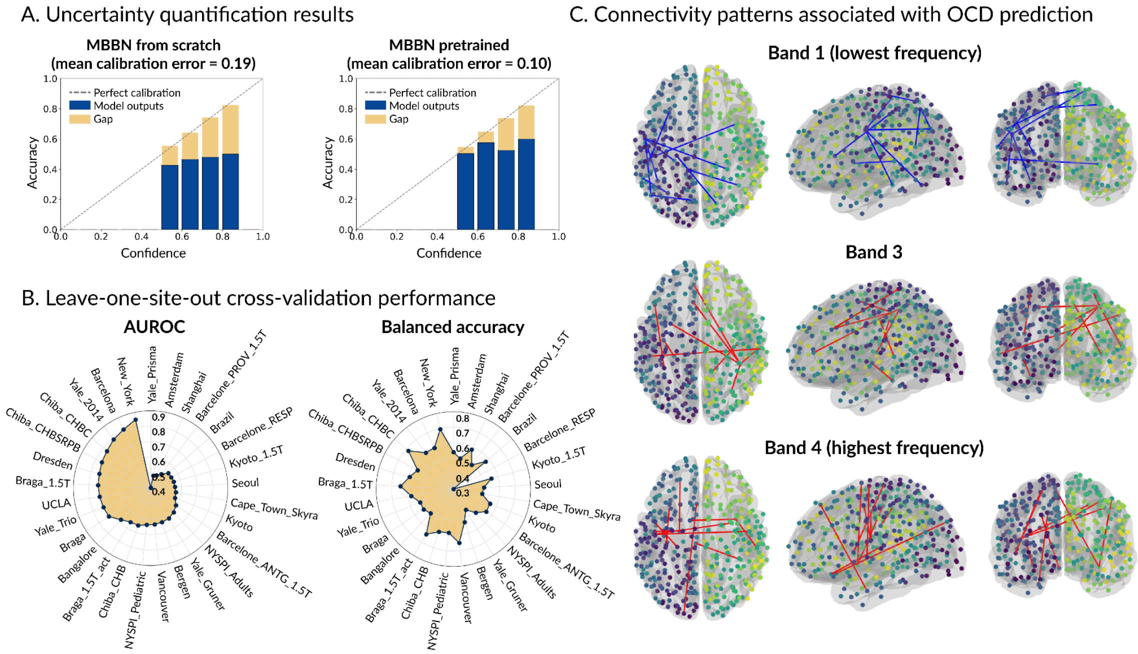

Figure 2: **(A)** Calibration error plots comparing predicted confidence with empirical accuracy in the Multi-Band Brain Net (MBBN) model. Pretraining on the large-scale UK Biobank data has reduced the mean calibration error from 0.19 to 0.10. **(B)** Leave-one-site-out cross-validation results showing AUROC and balanced accuracy for each test site using the pretrained MBBN model. AUROC ranged from .40 to .85, and balanced accuracy ranged from .30 to .70. **(C)** Self-attention-based functional connectivity patterns associated with OCD classification across frequency bands. The top 10 significant connections with the largest effect sizes are displayed in blue (lower in the OCD group) and red (higher in the OCD group). Band cutoffs differ among subjects. No significant connections were observed in band 2.

Table 1. **OCD classification performance of the Multi-Band Brain Net (MBBN) compared with baseline models.** Test AUROC and balanced accuracy are reported as mean ± standard deviation across three random data splits. The model with the highest test metrics is shown in **bold**. FC: functional connectivity, TS: time series.

Model (data form)	AUROC	Accuracy
XGBoost (FC)	.592 ± .018	.570 ± .012
SVM (FC)	.625 ± .022	.588 ± .007
BNT (FC)	.630 ± .022	.583 ± .036
BolT (TS)	.626 ± .025	.588 ± .021
vanilla BERT (TS)	.617 ± .040	.583 ± .026
MBBN from scratch (TS)	.635 ± .035	.598 ± .017
MBBN pretrained (TS)	**.666 ± .028**	**.637 ± .018**

Results and Conclusions

Our pretrained Transformer-based model achieved an AUROC of .666—an **absolute improvement of 7.4 percentage points** over XGBoost (AUROC of .592) and **4.9 points** over the next-best baseline (vanilla BERT at .617; Table 1). Accuracy likewise rose from .598 (from scratch) to .637 (pretrained). Uncertainty quantification showed an improved mean calibration error (0.10) relative to training from scratch (0.19), reflecting more reliable confidence estimates (Fig. 2A). LOSO-CV revealed site-level variability, with AUROC ranging from .40 to .85 and balanced accuracy from .30 to .70 (Fig. 2B). Self-attention analyses exposed **reduced sensorimotor network connectivity** and diminished links between **dorsal attention** and **temporoparietal** networks in OCD, as in previous findings (Fig. 2C; Bruin et al., 2023).

Despite residual cross-site heterogeneity, these results underscore that **large-scale pretraining** substantially enhances model generalizability and diagnostic accuracy for OCD, offering a promising path toward clinical utility.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1C1C1006503, RS-2023-00266787, RS-2023-00265406, RS-2024-00421268, RS-2024-00342301, RS-2024-00435727, NRF-2021M3E5D2A01022515), by Creative-Pioneering Researchers Program through Seoul National University(No. 200-20240057, 200-20240135), by Semi-Supervised Learning Research Grant by SAMSUNG(No.A0342-20220009), by Identify the network of brain preparation steps for concentration Research Grant by LooxidLabs(No.339-20230001), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] by the MSIT(Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program(RS-2024-00421268) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), by the National Supercomputing Center with supercomputing resources including technical support(KSC-2023-CRE-0568), by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A3A2A02090597), by the Korea Health Industry Development Institute (KHIDI), and by the Ministry of Health and Welfare, Republic of Korea (HR22C1605), by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea) & Gwangju Metropolitan City and by KBRI basic research program through Korea Brain Research Institute funded by Ministry of Science and ICT(25-BR-05-01).

References

Bae, S., Kwon, J., Yoo, S., & Cha, J. (2025). Spatiotemporal Learning of Brain Dynamics from fMRI Using Frequency-Specific Multi-Band Attention for Cognitive and Psychiatric Applications (No. arXiv:2503.23394). [Preprint] arXiv. https://doi.org/10.48550/arXiv.2503.23394

Bruin, W. B., Abe, Y., Alonso, P., Anticevic, A., Backhausen, L. L., Balachander, S., Bargallo, N., Batistuzzo, M. C., Benedetti, F., Bertolin Triquell, S., Brem, S., Calesella, F., Couto, B., Denys, D. A. J. P., Echevarria, M. A. N., Eng, G. K., Ferreira, S., Feusner, J. D., Grazioplene, R. G., … van Wingen, G. A. (2023). The functional connectome in obsessive-compulsive disorder: Resting-state mega-analysis and machine learning classification for the ENIGMA-OCD consortium. *Molecular Psychiatry*, *28*(10), 4307–4319. https://doi.org/10.1038/s41380-023-02077-0

Bruin, W., Denys, D., & van Wingen, G. (2019). Diagnostic neuroimaging markers of obsessive-compulsive disorder: Initial evidence from structural and functional MRI studies. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *91*, 49–59. https://doi.org/10.1016/j.pnpbp.2018.08.005

Dragomiretskiy, K., & Zosso, D. (2014). Variational Mode Decomposition. *IEEE Transactions on Signal Processing*, *62*(3), 531–544. IEEE Transactions on Signal Processing. https://doi.org/10.1109/TSP.2013.2288675

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.