Leveraging Generic Time Series Foundation Models for EEG Classification

Théo Gnassounou*
Université Paris-Saclay, Inria, CEA

Yessin Moakher*
Huawei Noah's Ark Lab

Shifeng Xie
Huawei Noah's Ark Lab

Vasilii Feofanov
Huawei Noah's Ark Lab

Ievgen Redko
Huawei Noah's Ark Lab

Abstract

Foundation models for time series are emerging as powerful general-purpose backbones, yet their potential for domain-specific biomedical signals such as electroencephalography (EEG) remains rather unexplored. In this work, we investigate the applicability a recently proposed time series classification foundation model, to a different EEG tasks such as motor imagery classification and sleep stage prediction. We test two pretraining regimes: (a) pretraining on heterogeneous real-world time series from multiple domains, and (b) pretraining on purely synthetic data. We find that both variants yield strong performance, consistently outperforming EEGNet, a widely used convolutional baseline, and CBraMod, the most recent EEG-specific foundation model. These results suggest that generalist time series foundation models, even when pretrained on data of non-neural origin or on synthetic signals, can transfer effectively to EEG. Our findings highlight the promise of leveraging cross-domain pretrained models for brain signal analysis, suggesting that EEG may benefit from advances in the broader time series literature.

1 Introduction

Electroencephalography (EEG) is a widely used non-invasive technique for monitoring brain activity, with applications ranging from clinical diagnostics to brain–computer interfaces (BCI). A central challenge in EEG analysis is classification, which underpins tasks such as motor imagery for BCI control (Barachant et al., 2012), sleep staging (Chambon et al., 2018), and emotion recognition (Li et al., 2022). Despite their promise, these applications face significant barriers: EEG datasets are typically small, fragmented across institutions, and/or difficult to share due to privacy concerns. Furthermore, EEG signals exhibit strong variability across subjects and sessions, which makes generalization to unseen individuals especially difficult. This scarcity and variability limit the effectiveness of deep learning models such as CNNs (Lawhern et al., 2018; Chambon et al., 2018), LSTMs (Phan et al., 2019), and transformers (Phan et al., 2022; Guo et al., 2024).

In parallel, machine learning has been transformed by the rise of foundation models (Bommasani et al., 2021). In computer vision (Dosovitskiy et al., 2021) and natural language processing (Achiam et al., 2023), large-scale pretraining on diverse data has enabled models to generalize across tasks, reducing the need for task-specific architectures and large labeled datasets. Inspired by this success, time series foundation models (TSFMs) have recently emerged. Some focus on forecasting (Ansari et al., 2024; Auer et al., 2025), others on classification (Feofanov et al., 2025), also with some attempts to unify multiple time series tasks (Goswami et al., 2024). Interestingly, both real-world (Feofanov et al., 2025) and synthetic data (Xie et al., 2025) have been shown effective for pretraining such models.

*equal contribution

NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT^{2}S).

Table 1: Model comparison.

Model	Type	Domain	Size	Pretraining Type	Multivariate Pretraining
EEGNet	Tailored	EEG	<0.01M	NaN	NaN
CBraMod	Foundation Model	EEG	4M	Reconstruction	Yes
Mantis	Foundation Model	Generic	8M	Contrastive	No

In EEG specifically, efforts to build foundation models are more recent. CBraMod (Wang et al., 2025) introduced a masked-reconstruction approach pretrained on the large-scale TUEG corpus (Obeid and Picone, 2016), showing encouraging results for different BCI tasks. Yet, its evaluation remains limited in scope, and questions persist about whether EEG-specific pretraining is necessary, or whether general-purpose TSFMs can transfer effectively to EEG.

This paper takes a step forward to address the aforementioned questions. We investigate the applicability of Mantis (Feofanov et al., 2025), the most recent time series classification foundation model pretrained either on heterogeneous time series datasets or synthetic data, to EEG signals. Across benchmarks for sleep staging and motor imagery classification, we find that Mantis achieves strong transfer performance, generally outperforming both EEGNet (Lawhern et al., 2018), a widely used baseline, and CBraMod, the most recent EEG-specific foundation model. This result, at the same time, gives a high promise on developing general-purpose TSFMs and highlights a large room for improvement in brain signal analysis. This finding suggests opportunities for leveraging cross-domain TSFMs in brain signal analysis as well as reveals current limitations of EEG-specific foundation models.

2 Methodology

Time series classification foundation model is an encoder $F : \mathbb{R}^{C\times T} \to \mathbb{R}^{Q}$ that projects any signal $\mathbf{x} \in \mathbb{R}^{C\times T}$ with C channels and sequence length T to a discriminative hidden space $\mathbb{R}^{Q}$. During pretraining, we observe an unlabeled pretraining set X_0 that is sufficiently large for learning rich embeddings that generalize well across different tasks. During fine-tuning, we observe a supervised downstream task with observations $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$. We append a classification head $h : \mathbb{R}^{Q} \to \mathbb{R}^{K}$ and fine-tune $h \circ F$ by minimizing the cross-entropy loss. In this work, we consider two foundation models, which we briefly present below, and the summary can be found in Table 1.

The first foundation model is CBraMod (Wang et al., 2025), recently proposed for EEG data. It is a masked autoencoder (He et al., 2022) focused on correct reconstruction of missing patches. The model has been pretrained on the TUEG dataset (Obeid and Picone, 2016) with 1,109,545 EEG samples after pre-processing. One of the main features of the model is that it is pretrained directly on the multivariate signals. This is achieved by the proposed criss-cross transformer that mixes time-wise and channel-wise attention modules. More implementation details can be found in Appendix A.1.

Second, we consider Mantis (Feofanov et al., 2025), a foundation model designed for general-purpose time series classification. In contrast to CBraMod, Mantis is pretrained using contrastive learning, pushing different augmentations of a single time series to lie close in the representation space. Originally, they have pretrained Mantis on a mix of different real-world time series datasets (1.8 millions samples in total, Lin et al.,2024), with a small portion of EEG data. Recently, Xie et al. (2025) has shown that Mantis achieves the same performance by being pretrained on a purely synthetic dataset generated by the CauKer algorithm (1 million samples). In our experiment, we will consider both these checkpoints. We give more details in Appendix A.2.

It is worth mentioning that in our preliminary experiments, we have found that freezing the encoder for EEG data leads to a huge decrease in performance, so fine-tuning is necessary in this context. This is why we have not considered MOMENT (Goswami et al., 2024), which is very difficult to fine-tune due to its large model size compared to CBraMod and Mantis. In our experiments, we compare the two foundation models with EEGNet (Lawhern et al., 2018), a classical CNN architecture specifically designed for EEG signals.

3 Experimental Results

3.1 Setup

We conduct two different sets of experiments to evaluate Mantis on EEG data. First, we follow the experimental setup used in CBraMod (Wang et al., 2025) and concentrate on motor imagery classification. Second, we perform a comprehensive study on 8 sleep stage prediction datasets following Perslev et al. (2021) and Gnassounou et al. (2025). While in the first case we extract the results of CBraMod and EEGNet from Wang et al. (2025), in the second case, we fine-tune these models, so we can test the adaptability of CBraMod to new EEG tasks.

BCI dataset In the first experiments we use two dataset of Brain Computer Interface for Motor Imagery classification. PhysioNet-MI (Schalk et al., 2004) comprises 109 subjects with 64 channels with a sampling rate of 160 Hz. This dataset includes four different motor imagery classes: left hand, right hand, both hands, and both feet. SHU-MI (Ma et al., 2022) comprises 25 subjects with 32 EEG channels sample at 250Hz. This dataset covers binary motor imagery with the right hand and the left hand. Each dataset is resampled at 200 Hz. For PhysioNet-MI, subjects 1–70, 71–89, and 90–109 are used for training, validation, and testing, respectively. For SHU-MI, subjects 1-15, 16-20, 21-25 for training, validation, and testing, respectively.

Sleep datasets In the experiments, we used 8 sleep staging datasets. ABC (Jessie P. et al., 2018), CCSHS (Rosen et al., 2003), CFS (Redline et al., 1995), HPAP (Rosen et al., 2012), MROS (Blackwell et al., 2011), SHHS (Zhang et al., 2018a), CHAT (Marcus et al., 2013), and SOF (Spira et al., 2008) are publicly available sleep datasets with restricted access from National Sleep Research Resource (NSRR) (Zhang et al., 2018b). PhysioNet (Goldberger et al., 2000) and MASS (O'Reilly et al., 2014) are two other datasets publicly available. These datasets are recordings of one night of sleep of different patients. Every 30 s epoch of sleep is labeled with one of the five sleep stages: Wake, N1, N2, N3, and REM. Datasets are split by subjects into training, validation, and test sets (60%/20%/20%). More details about the pre-processing is given in the Appendix.

Architecture setup For fine-tuning, Mantis use a linear layer with pre-LayerNorm as a classfication head. In the CBraMod's paper, they tune the head for each task, while in our experiments, we fixed it as a 3-layer MLP with ELU and dropout. For the BCI experiments, models were trained for a maximum of 20 epochs using the AdamW optimizer with a batch size of 64 and a weight decay of 0.01. We set the initial learning rate to 1×10^{-4} for the Physionet-MI dataset and 5×10^{-4} for the SHU-MI dataset. The learning rate was managed by a cosine scheduler with a warmup period over the first 20% of training steps. We applied gradient clipping at a norm of 1.0 and utilized an early stopping mechanism with a patience of 3 epochs to prevent overfitting.

For sleep staging, models were trained for a maximum of 50 epochs using the AdamW optimizer with a batch size of 64 and a weight decay of 0.01. Training is monitored with early stopping on the validation set, using a patience of 5 epochs. The learning rate is set to 1×10^{-4} for Mantis and CBraMod, and 1×10^{-3} for EEGNet. To account for limited resources, we impose a maximum training time of 5 hours. A value of NaN indicates that the time limit was reached.

3.2 Results on Brain Computer Interface

We evaluated Mantis on a BCI motor imagery task, using the CBraMod setup for comparison (Table 2). On the Physionet-MI dataset, Mantis achieves performance competitive with CBraMod, a specialized model for EEG that already surpasses classical CNNs like EEGNet by 6%. This similar result is particularly noteworthy as Mantis was pretrained with minimal EEG data. More importantly, Mantis significantly outperforms the baseline on the SHU-MI dataset, achieving a 72.15% F1-score versus CBraMod's 69.88%, when Mantis was pretrained only on synthetic data (Xie et al., 2025). These results demonstrate that its architecture can achieve state-of-the-art performance on brain signals without extensive domain-specific pretraining.

Mantis's performance is even more compelling given its approach to modeling spatial correlations, which are critical for BCI tasks (Barachant et al., 2012). Unlike CBraMod, which relies on multivariate pretraining, Mantis processes channels univariately and only models their inter-dependencies at the final classification layer. The fact that Mantis still outperforms the multivariate approach suggests its

architecture offers a more efficient method for preserving and leveraging spatial information in EEG signals.

3.3 Results on Sleep Staging

We then evaluate the models on sleep staging, a task characterized by a low number of EEG channels (typically 1-7) (Supratak et al., 2017; Gnassounou et al., 2024; Wang et al., 2025), contrasting with the highly multivariate signals CBraMod was designed for. As shown in Table 3 for a 2-channel setup, both foundation models surpass the EEGNet baseline. Crucially, Mantis consistently outperforms CBraMod across all pretraining configurations (real and synthetic data), with performance gains ranging from 0.3% on CCSHS to nearly 3% on the Mass dataset. This suggests that in scenarios with limited spatial information, CBraMod's multivariate architecture is less effective, whereas Mantis's more general, channel-independent design holds a distinct advantage.

Additionally, our results confirm the value of pretraining, as starting from a checkpoint yields better performance than random initialization. Pretraining also enhances training stability and efficiency, preventing runtime errors (denoted by NaN for runs exceeding 5 hours) that occurred when training from scratch. However, the performance gains are modest, indicating substantial room for improvement in future pretraining strategies for EEG data.

Table 2: **Three different scores for BCI over two datasets averaged over 3 seeds.** EEGNet and CBraMod results are from (Wang et al., 2025). For Mantis, we report random initialization (Random), pretraining on real dataset (Real Pretrain) and synthetic pretraining on data generated by CauKer (Xie et al., 2025) (Synth Pretrain).

Dataset	Metric	EEGNet	CBraMod	Mantis		
				Random	Real Pretrain	Synth Pretrain
PhysioNet-MI	Balanced Acc	$58.14_{\pm 1.25}$	$64.17_{\pm 0.90}$	$60.76_{\pm 1.10}$	$\mathbf{64.43_{\pm 1.50}}$	$61.90_{\pm 2.01}$
	Cohen's Kappa	$44.68_{\pm 1.20}$	$\mathbf{52.22_{\pm 1.70}}$	$47.70_{\pm 1.46}$	$52.13_{\pm 1.87}$	$49.20_{\pm 3.35}$
	Weighted F1	$57.96_{\pm 1.15}$	$64.27_{\pm 1.00}$	$60.44_{\pm 1.27}$	$\mathbf{64.34_{\pm 1.63}}$	$61.95_{\pm 2.52}$
SHU-MI	Balanced Acc	$58.89_{\pm 1.77}$	$63.70_{\pm 1.50}$	$60.70_{\pm 1.90}$	$63.00_{\pm 1.37}$	$\mathbf{65.5_{\pm 4.3}}$
	AUROC	$63.11_{\pm 1.42}$	$\mathbf{71.39_{\pm 0.90}}$	$68.10_{\pm 0.07}$	$69.46_{\pm 1.18}$	$70.90_{\pm 4.1}$
	AUC-PR	$62.83_{\pm 1.52}$	$69.88_{\pm 0.07}$	$70.00_{\pm 0.09}$	$70.55_{\pm 2.00}$	$\mathbf{72.15_{\pm 3.8}}$

Table 3: **Weighted F1 score for sleep staging over eight datasets averaged over 3 random seeds.** Random and Synth Pretrain settings are as described in Table 2.

Dataset	EEGNet	CBraMod		Mantis		
		Random	EEG Pretrain	Random	Real Pretrain	Synth Pretrain
ABC	$67.94_{\pm 6.52}$	$70.61_{\pm 3.29}$	$74.90_{\pm 4.89}$	$72.82_{\pm 3.89}$	$75.50_{\pm 5.62}$	$\mathbf{75.74_{\pm 4.32}}$
CCSHS	$83.13_{\pm 0.10}$	$87.01_{\pm 0.27}$	$88.04_{\pm 0.59}$	$88.55_{\pm 0.39}$	$\mathbf{88.85_{\pm 0.48}}$	$88.80_{\pm 0.30}$
CFS	$78.60_{\pm 1.31}$	$83.48_{\pm 0.23}$	$84.30_{\pm 0.08}$	$84.96_{\pm 0.43}$	$\mathbf{85.35_{\pm 0.35}}$	$85.06_{\pm 0.75}$
CHAT	$78.91_{\pm 0.16}$	$84.11_{\pm 0.81}$	$85.01_{\pm 0.42}$	NaN	$\mathbf{85.94_{\pm 0.18}}$	$85.72_{\pm 0.29}$
HOMEPAP	$69.43_{\pm 0.08}$	$70.37_{\pm 1.90}$	$72.56_{\pm 2.35}$	$71.26_{\pm 1.93}$	$73.14_{\pm 2.09}$	$\mathbf{73.53_{\pm 2.00}}$
MASS	$79.85_{\pm 1.27}$	$77.40_{\pm 2.18}$	$81.12_{\pm 2.27}$	$79.06_{\pm 1.89}$	$\mathbf{84.09_{\pm 0.85}}$	$82.49_{\pm 1.22}$
PhysioNet	$75.73_{\pm 0.38}$	$77.19_{\pm 0.94}$	$78.97_{\pm 0.43}$	$77.98_{\pm 0.89}$	$\mathbf{79.82_{\pm 1.63}}$	$78.83_{\pm 1.60}$
SOF	$78.74_{\pm 1.81}$	$82.61_{\pm 0.35}$	$83.39_{\pm 0.67}$	$83.70_{\pm 1.01}$	$\mathbf{84.69_{\pm 0.73}}$	$84.31_{\pm 0.57}$

4 Conclusion / Open Challenges

While promising, the current findings on Mantis's superior performance for EEG analysis suggest a significant step forward in applying foundation models to neural data. Its ability to outperform a specialized EEG model highlights the potential of generalist architectures. Future work should focus on extending these experiments to include a wider range of BCI datasets and new tasks, like emotion recognition, to fully validate Mantis's generalizability. Furthermore, addressing the challenge of zero-shot learning on EEG data, perhaps through specialized normalization techniques such as Monge alignment Gnassounou et al. (2025), could unlock new avenues for leveraging these powerful models without extensive training.

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*.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. (2024). Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Appelhoff, S., Sanderson, M., Brooks, T. L., van Vliet, M., Quentin, R., Holdgraf, C., Chaumon, M., Mikulan, E., Tavabi, K., Höchenberger, R., Welke, D., Brunner, C., Rockhill, A. P., Larson, E., Gramfort, A., and Jas, M. (2019). MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis. *Journal of Open Source Software*, 4(44):1896.

Auer, A., Podest, P., Klotz, D., Böck, S., Klambauer, G., and Hochreiter, S. (2025). Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *arXiv preprint arXiv:2505.23719*.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2012). Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928.

Blackwell, T., Yaffe, K., Ancoli-Israel, S., Redline, S., Ensrud, K. E., Stefanick, M. L., Laffan, A., Stone, K. L., and Osteoporotic Fractures in Men Study Group (2011). Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the Osteoporotic Fractures in Men Sleep Study. *Journal of the American Geriatrics Society*, 59(12):2217–2225.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., and Gramfort, A. (2018). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feofanov, V., Wen, S., Alonso, M., Ilbert, R., Guo, H., Tiomoko, M., Pan, L., Zhang, J., and Redko, I. (2025). Mantis: Lightweight calibrated foundation model for user-friendly time series classification. *arXiv preprint arXiv:2502.15637*.

Gnassounou, T., Collas, A., Flamary, R., and Gramfort, A. (2025). Psdnorm: Test-time temporal normalization for deep learning in sleep staging. *arXiv preprint arXiv:2503.04582*.

Gnassounou, T., Collas, A., Flamary, R., Lounici, K., and Gramfort, A. (2024). Multi-source and test-time domain adaptation on multivariate signals using spatio-temporal monge alignment.

Goldberger, A., Amaral, L., Glass, L., Havlin, S., Hausdorg, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K., Stanley, H., and Physiobank, P. (2000). Components of a new research resource for complex physiologic signals. *PhysioNet*, 101.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. (2024). MOMENT: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(267):1–13.

Guo, Y., Nowakowski, M., and Dai, W. (2024). Flexsleeptransformer: a transformer-based sleep staging model with flexible input channel configurations. *Scientific Reports*, 14.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Jessie P., B., Ali, T., Michael, R., Wei, W., Robert, A., Atul, M., Robert L., O., Amit, A., Katherine, D., and Sanya R., P. (2018). Gastric Banding Surgery versus Continuous Positive Airway Pressure for Obstructive Sleep Apnea: A Randomized Controlled Trial. *American journal of respiratory and critical care medicine*, 197(8). Publisher: Am J Respir Crit Care Med.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013.

Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., Zhao, Z., Kumar, N., and Marttinen, P. (2022). Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*, 55(4):1–57.

Lin, C., Wen, X., Cao, W., Huang, C., Bian, J., Lin, S., and Wu, Z. (2024). NuTime: Numerically multi-scaled embedding for large- scale time-series pretraining. *Transactions on Machine Learning Research*.

Ma, J., Yang, B., Qiu, W., Li, Y., Gao, S., and Xia, X. (2022). A large eeg dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9.

Marcus, C. L., Moore, R. H., Rosen, C. L., Giordani, B., Garetz, S. L., Taylor, H. G., Mitchell, R. B., Amin, R., Katz, E. S., Arens, R., Paruthi, S., Muzumdar, H., Gozal, D., Thomas, N. H., Ware, J., Beebe, D., Snyder, K., Elden, L., Sprecher, R. C., Willging, P., Jones, D., Bent, J. P., Hoban, T., Chervin, R. D., Ellenberg, S. S., Redline, S., and Childhood Adenotonsillectomy Trial (CHAT) (2013). A randomized trial of adenotonsillectomy for childhood sleep apnea. *The New England Journal of Medicine*, 368(25):2366–2376.

Obeid, I. and Picone, J. (2016). The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10.

O'Reilly, C., Gosselin, N., and Carrier, J. (2014). Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research*, 23.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P., and Igel, C. (2021). U-Sleep: resilient high-frequency sleep staging. *npj Digital Medicine*, 4:72.

Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and Vos, M. D. (2019). SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. arXiv:1809.10932.

Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., and De Vos, M. (2022). Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering*, 69(8):2456–2467.

Redline, S., Tishler, P. V., Tosteson, T. D., Williamson, J., Kump, K., Browner, I., Ferrette, V., and Krejci, P. (1995). The familial aggregation of obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine*, 151(3 Pt 1):682–687.

Rosen, C. L., Auckley, D., Benca, R., Foldvary-Schaefer, N., Iber, C., Kapur, V., Rueschman, M., Zee, P., and Redline, S. (2012). A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the HomePAP study. *Sleep*, 35(6):757–767.

Rosen, C. L., Larkin, E. K., Kirchner, H. L., Emancipator, J. L., Bivins, S. F., Surovec, S. A., Martin, R. J., and Redline, S. (2003). Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: association with race and prematurity. *The Journal of Pediatrics*, 142(4):383–389.

Schalk, G., Mcfarland, D., Hinterberger, T., Birbaumer, N., and Wolpaw, J. (2004). Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Trans. Biomed. Eng.*, 51:1034–.

Spira, A. P., Blackwell, T., Stone, K. L., Redline, S., Cauley, J. A., Ancoli-Israel, S., and Yaffe, K. (2008). Sleep-disordered breathing and cognition in older women. *Journal of the American Geriatrics Society*, 56(1):45–50.

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y. L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S. C., Kim, T. W., Pizza, F., Plazzi, G., Vandi, S., Antelmi, E., Perrin, D., Kuna, S. T., Schweitzer, P. K., Kushida, C., Peppard, P. E., Sorensen, H. B. D., Jennum, P., and Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9(1).

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008.

Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., and Pan, G. (2025). CBramod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*.

Xie, S., Feofanov, V., Alonso, M., Odonnat, A., Zhang, J., Palpanas, T., and Redko, I. (2025). Cauker: classification time series foundation models can be pretrained on synthetic data only. *arXiv preprint arXiv:2508.02879*.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. (2018a). The national sleep research resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association*, pages 572–572.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. (2018b). The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association: JAMIA*, 25(10):1351–1358.

A Architectures

A.1 CBraMod

Given an EEG sample $\mathbf{x} \in \mathbb{R}^{C \times T}$, where C is the number of channels and T is the number of timestamps, CBraMod first partitions the time axis into non-overlapping windows of length t ($t = 200$ used in pretraining), producing

$$\mathbf{x} \mapsto \mathbf{X} \in \mathbb{R}^{C \times p \times t}, \quad p = \left\lfloor \frac{T}{t} \right\rfloor.$$

Each patch $\mathbf{x}_{i,j}$ (short time series from channel i and window j) is independently encoded via:

- a time-domain convolutional branch f_{conv} (3-layer 1D CNN),
- a frequency-domain branch $W_{\text{fft}} \cdot \text{FFT}(\cdot)$ (FFT + linear projection).

Formally,

$$\mathbf{e}^t_{i,j} = f_{\text{conv}}(\mathbf{x}_{i,j}) \in \mathbb{R}^{200}, \quad \mathbf{e}^f_{i,j} = W_{\text{fft}} \cdot \text{FFT}(\mathbf{x}_{i,j}) \in \mathbb{R}^{200}.$$

The embeddings are combined as

$$\mathbf{e}_{i,j} = \mathbf{e}^t_{i,j} + \mathbf{e}^f_{i,j}, \quad \mathbf{E} \in \mathbb{R}^{C \times p \times 200}.$$

Asymmetric Conditional Positional Encoding (ACPE) generates spatial-temporal offsets $\mathbf{e}^{pos}_{i,j} \in \mathbb{R}^{200}$, which are added to the patch embeddings:

$$\mathbf{o}_{i,j} = \mathbf{e}_{i,j} + \mathbf{e}^{pos}_{i,j}, \quad \mathbf{O} \in \mathbb{R}^{C \times p \times 200}.$$

The embeddings $\mathbf{O}$ are processed through $L = 12$ criss-cross transformer blocks, each with parallel spatial and temporal attention:

$$\begin{aligned} \mathbf{O}_j &\in \mathbb{R}^{C \times 200}, & \text{S-Attn}(\mathbf{O}_j) &= \text{Attention}(\mathbf{O}_j W^Q, \mathbf{O}_j W^K, \mathbf{O}_j W^V), \\ \mathbf{O}_i &\in \mathbb{R}^{p \times 200}, & \text{T-Attn}(\mathbf{O}_i) &= \text{Attention}(\mathbf{O}_i W^Q, \mathbf{O}_i W^K, \mathbf{O}_i W^V). \end{aligned}$$

Both attentions use $K = 8$ heads with hidden dimension $d = 200$, and the concatenated outputs yield

$$\mathbf{E}_r \in \mathbb{R}^{C \times p \times 200}.$$

Pretraining For masked autoencoding, each representation is projected back to the time domain:

$$\hat{\mathbf{x}}_{i,j} = W_r \mathbf{e}^r_{i,j} \in \mathbb{R}^t, \quad \hat{\mathbf{X}} \in \mathbb{R}^{C \times p \times t}.$$

When pretraining, 50% of the patches are masked. The reconstruction loss is applied only to masked patches:

$$\mathcal{L}_{\text{MAE}} = \|\hat{\mathbf{X}}_M - \mathbf{X}_M\|_2^2.$$

A.2 Mantis

Given a time series sample $\mathbf{x} \in \mathbb{R}^{C \times T}$, Mantis first resizes each channel to a fixed length t multiple of 32 ($t = 512$ used in pretraining) via interpolation and applies instance-level standardization (per-channel mean and variance over time). For channel i, let $\mathbf{x}^{(i)} \in \mathbb{R}^t$ denote the normalized series.

For each channel, a single 1D convolution layer (output width 256) followed by mean pooling produces 32 base patches. The same pipeline applied to the first difference $\Delta\mathbf{x}^{(i)}$ yields 32 differential patches. From the *raw* (pre-normalization) series, per-patch statistics (mean μ_j and standard deviation σ_j) are computed and encoded via a scalar encoder. Concatenating the three feature parts and projecting yields the final tokens:

$$\begin{aligned} \mathbf{c}_j &= \text{MeanPool}\big(\text{Conv}(\mathbf{x}^{(i)})\big)_j \in \mathbb{R}^{256}, \\ \mathbf{c}^\Delta_j &= \text{MeanPool}\big(\text{Conv}(\Delta\mathbf{x}^{(i)})\big)_j \in \mathbb{R}^{256}, \\ \mathbf{s}_j &= \text{ScalarEnc}(\mu_j, \sigma_j) \in \mathbb{R}^{64}, \\ \mathbf{t}_j &= \text{LayerNorm}\Big(W_{\text{proj}}\,[\,\mathbf{c}_j; \mathbf{c}^\Delta_j; \mathbf{s}_j\,]\Big) \in \mathbb{R}^{256}, \quad j = 1, \ldots, 32, \end{aligned}$$

so that

$$T^{(i)} = [\mathbf{t}_1, \ldots, \mathbf{t}_{32}] \in \mathbb{R}^{32\times 256}.$$

A learnable class token $\mathbf{t}_{\text{cls}}$ is prepended, and sinusoidal positional encodings P are added:

$$\tilde{T}^{(i)} = [\mathbf{t}_{\text{cls}}; T^{(i)}] + P \in \mathbb{R}^{33\times 256}.$$

The sequence $\tilde{T}^{(i)}$ is processed through $L = 6$ Transformer encoder blocks (each with $H = 8$ heads; dropout 0.1 during pretraining), and the class embedding is taken as the channel descriptor:

$$\mathbf{z}^{(i)} = \mathrm{ViT}_L(\tilde{T}^{(i)})_{\text{cls}} \in \mathbb{R}^{256}.$$

All channels are encoded independently and concatenated:

$$\mathbf{z} = \mathrm{concat}\big(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(C)}\big) \in \mathbb{R}^{256\cdot C}.$$

Pretraining Mantis was pretrained using a contrastive loss. Let $\mathbf{x}$ and $\mathbf{x}'$ be two augmented views of the same original time series, and let their encoded representations be

$$\mathbf{z} = \mathrm{Mantis}(\mathbf{x}), \qquad \mathbf{z}' = \mathrm{Mantis}(\mathbf{x}') \in \mathbb{R}^{256\cdot C}.$$

We define the cosine similarity between two vectors as

$$\mathrm{scos}(\mathbf{a}, \mathbf{b}) = \Big\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2}, \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \Big\rangle.$$

For a batch of N samples, the contrastive (InfoNCE) loss for the i-th sample is

$$\mathcal{L}_i = -\log \frac{\exp\big(\mathrm{scos}(\mathbf{z}_i, \mathbf{z}'_i)/\tau\big)}{\sum_{j=1}^{N} \exp\big(\mathrm{scos}(\mathbf{z}_i, \mathbf{z}'_j)/\tau\big)},$$

where $\tau > 0$ is a temperature hyperparameter. The total loss is averaged over the batch:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i.$$

This training encourages embeddings of augmented views of the same sample to be close, while pushing apart embeddings of different samples, yielding representations that capture meaningful temporal features invariant to augmentations.

B Experimental Setup

Table 4: Datasets and number of subjects for sleep datasets.

Dataset	ABC	CCSHS	CFS	HPAP	PHYS	MASS	CHAT	SOF
Subjects	44	515	681	166	70	61	1230	434

Table 4 reports the number of subjects in the sleep datasets. As shown, the number of subjects varies widely, ranging from 44 in ABC to 1,230 in CHAT. For sleep dataset, we adopt a standard pre-processing step commonly used in sleep staging studies (Chambon et al., 2018; Stephansen et al., 2018). To ensure consistency across, we restrict the analysis to two bipolar EEG channels. For the NSRR datasets, we select C3-A2 and C4-A1, while for Physionet and MASS, only Fpz-Cz and Pz-Oz are available and thus used. All EEG signals are low-pass filtered at 30 Hz and resampled to 100 Hz. For CBraMod, the data are resampled to 200Hz and split into 1s patches. Data extraction and preprocessing are performed with MNE-BIDS (Appelhoff et al., 2019) and MNE-Python (Gramfort et al., 2013).