Are foundation models useful feature extractors for electroencephalography analysis?

Özgün Turgut[1, 2] **Felix S. Bott**[2, 3] **Markus Ploner**[2, 3, 4] **Daniel Rueckert**[1, 2, 5, 6]

[1]School of Computation, Information and Technology, Technical University of Munich, Germany
[2]School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Germany
[3]Department of Neurology, Technical University of Munich, Germany
[4]Center for Interdisciplinary Pain Medicine, Technical University of Munich, Germany
[5]Munich Center for Machine Learning, Munich, Germany
[6]Department of Computing, Imperial College London, United Kingdom
{oezguen.turgut, daniel.rueckert}@tum.de

Abstract

The success of foundation models in natural language processing and computer vision has motivated similar approaches in time series analysis. While foundational time series models have proven beneficial on a variety of tasks, their effectiveness in medical applications with limited data remains underexplored. In this work, we investigate this question in the context of electroencephalography (EEG) by evaluating general-purpose time series models on age prediction, seizure detection, and classification of clinically relevant EEG events. We compare their diagnostic performance against specialised EEG models and assess the quality of the extracted features. The results show that general-purpose models are competitive and capture features useful to localising demographic and disease-related biomarkers. These findings indicate that foundational time series models can reduce the reliance on large task-specific datasets and models, making them valuable in clinical practice.

1 Introduction

Recent breakthroughs in natural language processing and computer vision have shown the effectiveness of foundation models on a wide range of tasks. Inspired by this success, a growing number of works has focused on developing similar models for time series analysis [3, 6, 8, 11, 17, 23, 26, 28, 31]. While most of the foundational time series models are designed for only a single task such as forecasting [3, 17, 23, 28], recent works [6, 8, 26] have introduced general-purpose models that are effective on diverse tasks, including regression, classification, and forecasting. This raises the questions of (1) whether and (2) how *general-purpose models* can be *translated into a medical context* to benefit clinical applications with limited data availability.

One relevant clinical application is electroencephalography (EEG), a widely accessible and cost-effective tool for measuring electrical brain activity across frequencies ranging from 0.5 to $100\,\text{Hz}$, typically grouped into unified bands: delta (δ: 0.5–4 Hz), theta (θ: 4–8 Hz), alpha (α: 8–13 Hz), beta (β: 13–30 Hz), and gamma (γ: 30–100 Hz) [18]. Despite its accessibility, large-scale EEG datasets ($>10\,\text{k}$ samples) remain scarce, limiting the ability to learn generalisable EEG features. As a result, existing EEG models are typically task-specific, with architectures and training schemes tailored to applications such as age prediction [2, 5], seizure detection [1], or event type classification [9]. This task-specific nature leads to poor model generalisability and requires substantial effort to design and train new models for each use case. In contrast, clinicians might tackle diverse EEG use cases more

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

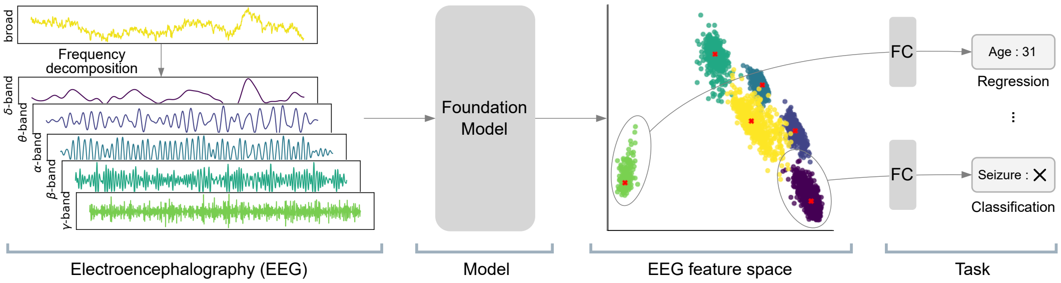


Figure 1: **Overview**. We study the ability of general-purpose time series models to extract meaningful electroencephalography (EEG) features for tasks such as age prediction or seizure detection.

Table 1: Overview of datasets used for regression and classification solely from EEG.

Dataset	**Task**	**# Samples**	**# Variates**	**# Time points**		**Frequency**
				per sample	*total*	
LEMON [2]	Age prediction	378	32	$30,000$	$11\,\mathrm{M}$	$250\,\mathrm{Hz}$
Epilepsy [1]	Seizure detection (Binary)	$11,500$	1	178	$2\,\mathrm{M}$	$174\,\mathrm{Hz}$
TUEV [9]	Event classification (Multi-class)	$112,237$	19	$1,000$	$112\,\mathrm{M}$	$200\,\mathrm{Hz}$

effectively using general-purpose models, given their general time series knowledge obtained through pre-training on large, heterogeneous datasets ($>100\,\mathrm{k}$ samples).

To this end, we systematically investigate the applicability of general-purpose models to EEG analysis (see Figure 1). In our study, we (1) compare their diagnostic performance against specialised EEG models across three public datasets, (2) evaluate the necessity of domain adaptation, and (3) analyse their ability to localise demographic and disease-related information.

2 Materials & Methods

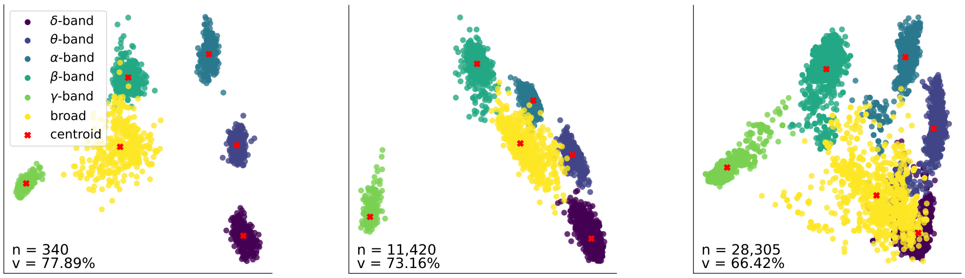
2.1 General-Purpose Models & Domain Adaptation Strategies

Our study includes three general-purpose models, namely `MOMENT` (●) [8], `UniTS` (●) [6], and `OTiS` (●) [26]. These models are pre-trained on large, heterogeneous time series corpora to learn general time series features. For instance, the most recent `OTiS` was pre-trained on $640,187$ time series samples from 8 domains, including $400,000$ ECG ($62.48\,\%$), $203,340$ weather ($31.76\,\%$), $19,614$ audio ($3.06\,\%$), $13,640$ engineering ($2.13\,\%$), $3,367$ EEG ($0.53\,\% = 0.42\,\%$ resting-state EEG $+$ $0.11\,\%$ emotion recognition EEG; totalling 125 recording hours), 115 economics ($0.02\,\%$), and 111 banking ($0.02\,\%$) samples. Moreover, to ensure a fair comparison with specialised EEG models, free from architectural, training, and scaling biases, we additionally pre-train `OTiS` from sratch exclusively on the $3,367$ EEG samples following [26] and include this specialised variant as $\texttt{OTiS}_{\text{EEG}}$ (●).

We evaluate three domain adaptation strategies. For **zero-shot** (ZS), the model is frozen after pre-training and evaluated without any fine-tuning. Its output tokens are averaged to obtain a global representation. Class logits are computed via cosine similarity between a test sample's representation and each class representation, i.e. the mean global representation of all training samples from a class. This adaptation strategy applies only for classification, while the following two also support regression. For **linear probing**, the model remains frozen while a randomly initialised linear layer is trained. For **fine-tuning** (FT), both the model and a randomly initialised linear layer are trained.

2.2 Datasets

Our extensive experiments include three publicly available datasets covering distinct tasks, as detailed in Table 1. **LEMON** [2] comprises resting-state EEG sampled at $250\,\mathrm{Hz}$ from healthy subjects aged $20-35$ years ($67\,\%$) and $59-77$ years ($33\,\%$). **Epilepsy** [1] includes single-channel EEG from healthy subjects at rest ($20\,\%$) and patients during epileptical seizures ($80\,\%$), sampled at $174\,\mathrm{Hz}$ and band-pass filtered between $0.5-40\,\mathrm{Hz}$. **TUEV** [9] is a large EEG corpus with patient recordings of



(a) LEMON. (b) Epilepsy. (c) TUEV.

Figure 2: PCA of zero-shot EEG features. By capturing distinct EEG features across frequency bands, general-purpose models such as `OTiS` may offer valuable benefits for clinical practice.

three clinically relevant events, including spike and sharp waves (SPSW: $2\,\%$), generalised periodic epileptiform discharges (GPED: $7.06\,\%$), and periodic lateralised epileptiform discharges (PLED: $12.58\,\%$), as well as three noise events, such as eye movement (EYEM: $1.16\,\%$), artifacts from equipment or the environment (ARTF: $7.79\,\%$), and background activity (BCKG: $69.41\,\%$).

2.3 Experimental Setup

2.3.1 Processing & Evaluation.

We follow established data processing, splitting, and evaluation protocols for age regression on LEMON [5] and classification on Epilepsy [33] and TUEV [31], reporting results across five random seeds for both linear probing and fine-tuning. To this end, we measure the coefficient of determination (R^2) for regression on LEMON and accuracy (ACC)/balanced accuracy (bACC) for classification on Epilepsy/TUEV. Regression and classification tasks are optimised using a mean squared error and cross-entropy loss, respectively. Training is performed using early stopping, and the model achieving the best validation performance is evaluated on the test set. Optimal hyperparameters are found through a grid search over the learning rate (3e-5, 1e-4, 3e-4, 1e-3, 3e-3), batch size (2^x, $x \in [2, 3, ..., 7]$), drop path (0.1, 0.2), layer decay (0.5, 0.75), weight decay (0.0, 0.1, 0.2), and label smoothing (0.0, 0.1, 0.2). All experiments are conducted on one NVIDIA RTX A6000-48GB GPU.

2.3.2 Baselines.

We benchmark the general-purpose models against 16 specialised EEG models (†), including 2 foundational EEG models (‡), and 4 statistical feature-based approaches (*). For age prediction, we compare against regression toward the mean (RTM; predictions equal the training data's mean age)*, handcrafted features* [5], the filterbank Riemann model* [21], the filterbank source model* [5], shallow ConvNet† [22], and deep ConvNet† [22]. For seizure detection, we include SimCLR† [25], TimesNet† [29], CoST† [27], TS2Vec† [32], TF-C† [33], Ti-MAE† [16], and SimMTM† [4], all pre-trained on SleepEEG [14] totalling 205 recording hours. For EEG event type classification, the baselines comprise ST-Transformer† [24], CNN-Transformer† [20], FFCL† [15], SPaRCNet† [12], ContraWR† [30], BIOT‡ ($3\,$M parameter, pre-trained on $13,000$ recording hours) [31], and LaBraM‡ ($370\,$M parameter, pre-trained on $2,500$ recording hours) [11].

3 Results & Discussion

General-purpose time series models capture distinct EEG features across frequency bands, as visualised in Figure 2, but their clinical utility remains unclear. To investigate this, we (1) compare the diagnostic performance of such models against specialised EEG models (Section 3.1), (2) evaluate whether they require domain adaptation to extract clinically relevant information (Section 3.2), and (3) analyse their potential to localise demographic and disease-related biomarkers (Section 3.3).

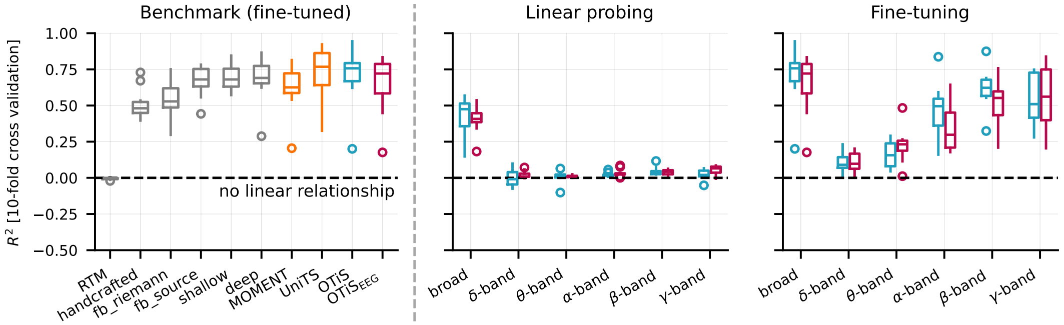


Figure 3: **LEMON**. (Left) General-purpose models (●,●) are competitive with specialised models (●,●). (Right) Fine-tuning is essential for clinical utility. Age effects are detected in higher frequencies.

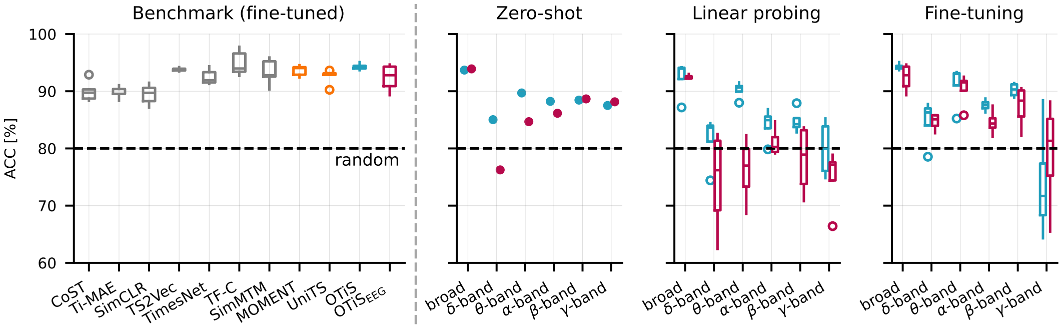


Figure 4: **Epilepsy**. (Left) General-purpose models (●,●) are competitive with specialised models (●,●). (Right) Domain adaptation through linear probing or fine-tuning is not required for seizure detection. Seizure-related information shows no clear frequency localisation.

3.1 Diagnostic Performance

We study whether general time series knowledge, obtained through pre-training on large and heterogeneous time series corpora, offers advantages for EEG analysis. To this end, we compare three general-purpose models - `MOMENT` (40 M) [8], `UniTS` (8 M) [6], and `OTiS` (7 M) [26] - against specialised EEG models trained solely on domain-specific data. Across benchmarks in age prediction, seizure detection, and EEG event classification, general-purpose models perform competitively, surpassing statistical baselines and in some cases even specialised models (see Figures 3, 4, and 5, left). Figure 5 suggests that general-purpose models might even capture clinically relevant information beyond what is achieved by specialised foundation models such as BIOT (3 M) [31]. Comparisons with the specialised $\texttt{OTiS}_{\text{EEG}}$ (7 M) further highlight the contribution of general time series pre-training, while ruling out differences in architecture and model size. Large-scale domain-specific pre-training can yield optimal performance, as indicated by LaBraM [11] in Figure 5, but such approaches are often limited by data availability. Overall, our findings suggest that general-purpose models can reduce reliance on large domain-specific datasets while retaining competitive diagnostic performance.

3.2 Domain Adaptation Strategies

We analyse whether general-purpose models can be applied to EEG analysis out-of-the-box or require domain adaptation. Therefore, we evaluate `OTiS` (●) and the specialised $\texttt{OTiS}_{\text{EEG}}$ (●) under zero-shot, linear probing, and fine-tuning settings (see *broad* in Figures 3, 4, and 5, right). Our experiments span datasets of varying scale: 11 M time points in LEMON, 2 M in Epilepsy, and 112 M in TUEV. We find that large domain-specific datasets, such as LEMON and TUEV, enable substantial improvements in feature quality through fine-tuning, whereas task-specific training on limited data, as in Epilepsy, risks performance loss through overfitting. Age prediction in LEMON proves particularly challenging, as evidenced by the linear probing results, and requires task-specific fine-tuning for clinical usability. For tasks where a visual interpretation of raw EEG is feasible, features of general-purpose models can be used out-of-the-box, as indicated by competitive zero-shot performances in Epilepsy and TUEV. In

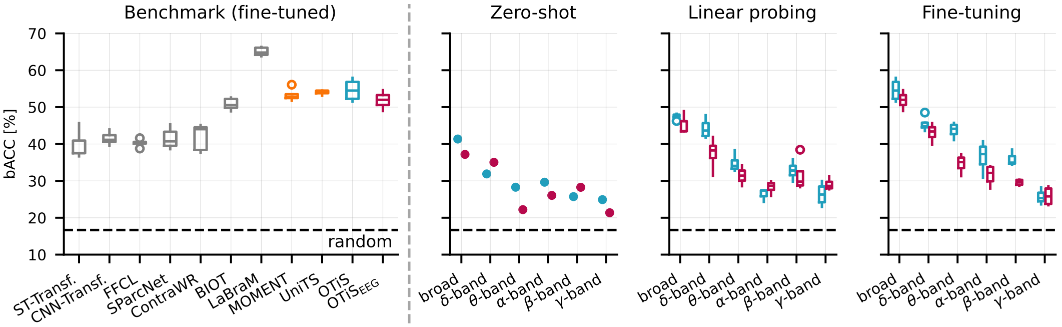
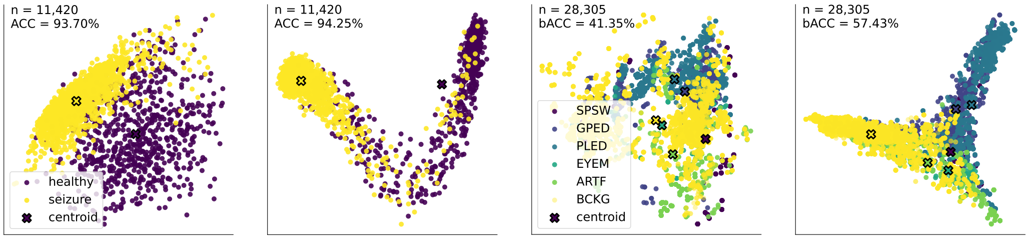


Figure 5: **TUEV**. (Left) General-purpose models (●,●) outperform nearly all specialised models (●,●). (Right) Zero-shot features are already expressive, but fine-tuning improves feature quality. Clinically relevant markers of seizure or acute neurological conditions are concentrated in lower frequencies.



(a) Epilepsy ZS.

(b) Epilepsy FT.

(c) TUEV ZS.

(d) TUEV FT.

Figure 6: PCA of EEG features extracted by `OTiS`. (a, b) Distinct features are captured for healthy subjects and patients, even in the zero-shot (ZS) setting. (c, d) The extraction of clinically relevant features is substantially enhanced with fine-tuning (FT).

particular, such readily available features are useful to detect spike and sharp waves (SPSW) indicative of epileptic seizures (see Figure 6a), or to distinguish SPSW from eye movement (EYEM) (see Figure 6c). Fine-tuning offers no advantages for simple tasks such as seizure detection (see Figures 6a and 6b) but is essential for complex tasks that require domain knowledge such as distinguishing EYEM from background activity (BCKG) (see Figures 6c and 6d). Finally, pre-training on large-scale heterogeneous time series data proves beneficial if domain-specific pre-training data is limited, as demonstrated by the comparison between `OTiS` and its specialised counterpart $\texttt{OTiS}_{\text{EEG}}$.

3.3 Biomarker Localisation

To evaluate whether general-purpose models enable the localisation of clinical biomarkers, we assess the expressiveness of EEG features across frequency bands (see Figures 3, 4, and 5, right). Optimal predictions are consistently obtained using broadband features, suggesting that they capture the full spectrum of clinically relevant information. Assessing predictions within individual frequency bands allows localisation of specific biomarkers: age-related information is encoded in higher frequencies (Figure 3), whereas ictal activity is concentrated in lower frequencies (Figure 5). These findings align with literature on age-related EEG biomarkers [10, 13] and epileptical markers [7, 19]. Band-specific analyses further reveal that data pre-processing heavily affects diagnostic performance: radical low-pass filtering of raw EEG at 40 Hz [1] reduces γ-band seizure detection to chance levels in Epilepsy (80% majority class; Figure 4). Interestingly, zero-shot features (Figure 2) already reflect these findings: broadband features align with features of the most informative bands (LEMON: β-band; TUEV: δ-band) and diverge from the ones of less informative bands (Epilepsy: γ-band).

4 Conclusion

In this study, we explore the utility of general-purpose time series models for electroencephalography (EEG) analysis. Through extensive benchmarking on age prediction, seizure detection, and EEG event type classification, we find that these models are competitive with specialised EEG models.

While general-purpose models can be used out-of-the-box for tasks where a visual interpretation of raw EEG is feasible (e.g. seizure detection), domain adaptation through fine-tuning becomes essential for tasks that require higher level of specialisation (e.g. event type classification). Furthermore, our experiments reveal that these models enable the localisation of demographic and disease-specific information through frequency band analysis. These findings indicate that foundation models can be useful in clinical routine, especially when domain-specific data is limited. Overall, we believe this work provides valuable guidance for integrating general-purpose models into clinical practice and motivates future exploration in other EEG tasks, such as sleep stage classification or motor imagery.

References

[1] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 2001.

[2] A. Babayan, M. Erbey, D. Kumral, J. D. Reinelt, A. M. Reiter, J. Röbbig, H. L. Schaare, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data*, 2019.

[3] A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

[4] J. Dong, H. Wu, H. Zhang, L. Zhang, J. Wang, and M. Long. Simmtm: A simple pre-training framework for masked time-series modeling. In *Advances in Neural Information Processing Systems*, 2024.

[5] D. A. Engemann, A. Mellot, R. Höchenberger, H. Banville, D. Sabbagh, L. Gemein, T. Ball, and A. Gramfort. A reusable benchmark of brain-age prediction from m/eeg resting-state signals. *Neuroimage*, 2022.

[6] S. Gao, T. Koker, O. Queen, T. Hartvigsen, T. Tsiligkaridis, and M. Zitnik. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 2024.

[7] N. Gaspard, L. Manganas, N. Rampal, O. A. Petroff, and L. J. Hirsch. Similarity of lateralized rhythmic delta activity to periodic lateralized epileptiform discharges in critically ill patients. *JAMA neurology*, 70(10):1288–1295, 2013.

[8] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski. Moment: A family of open time-series foundation models. *International Conference on Machine Learning*, 2024.

[9] A. Harati, M. Golmohammadi, S. Lopez, I. Obeid, and J. Picone. Improved eeg event classification using differential energy. In *IEEE Signal Processing in Medicine and Biology Symposium*, 2015.

[10] A. Hashemi, L. J. Pino, G. Moffat, K. J. Mathewson, C. Aimone, P. J. Bennett, L. A. Schmidt, and A. B. Sekuler. Characterizing population eeg dynamics throughout adulthood. *ENeuro*, 3 (6), 2016.

[11] W. Jiang, L. Zhao, and B. liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

[12] J. Jing, W. Ge, S. Hong, M. B. Fernandes, Z. Lin, C. Yang, S. An, A. F. Struck, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 2023.

[13] J.-H. Kang, J.-H. Bae, and Y.-J. Jeon. Age-related characteristics of resting-state electroencephalographic signals and the corresponding analytic approaches: a review. *Bioengineering*, 11(5):418, 2024.

[14] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 2000.

[15] H. Li, M. Ding, R. Zhang, and C. Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical signal processing and control*, 2022.

[16] Z. Li, Z. Rao, L. Pan, P. Wang, and Z. Xu. Ti-mae: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*, 2023.

[17] Y. Liu, G. Qin, X. Huang, J. Wang, and M. Long. Timer-XL: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations*, 2025.

[18] J. J. Newson and T. C. Thiagarajan. Eeg frequency bands in psychiatric disorders: a review of resting state studies. *Frontiers in human neuroscience*, 2019.

[19] D. Panet-Raymond and J. Gotman. Asymmetry in delta activity in patients with focal epilepsy. *Electroencephalography and clinical neurophysiology*, 75(6):474–481, 1990.

[20] W. Y. Peh, Y. Yao, and J. Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022.

[21] D. Sabbagh, P. Ablin, G. Varoquaux, A. Gramfort, and D. A. Engemann. Predictive regression modeling with meg/eeg: from source power to signals and cognitive states. *NeuroImage*, 2020.

[22] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, et al. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 2017.

[23] X. Shi, S. Wang, Y. Nie, D. Li, Z. Ye, Q. Wen, and M. Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. *International Conference on Learning Representations*, 2025.

[24] Y. Song, X. Jia, L. Yang, and L. Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

[25] C. I. Tang, I. Perez-Pozuelo, D. Spathis, and C. Mascolo. Exploring contrastive learning in human activity recognition for healthcare. *arXiv preprint arXiv:2011.11542*, 2020.

[26] Ö. Turgut, P. Müller, M. J. Menten, and D. Rueckert. Towards generalisable time series understanding across domains. *arXiv preprint arXiv:2410.07299*, 2025.

[27] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi. CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *International Conference on Learning Representations*, 2022.

[28] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. *International Conference on Machine Learning*, 2024.

[29] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2022.

[30] C. Yang, C. Xiao, M. B. Westover, J. Sun, et al. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2023.

[31] C. Yang, M. Westover, and J. Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 2024.

[32] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. Ts2vec: Towards universal representation of time series. In *AAAI Conference on Artificial Intelligence*, 2022.

[33] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This study presents an evaluation of time series foundation models for EEG analysis.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The study only covers classification and regression tasks.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work presents an empirical study.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This study fully supports reproducibility by detailing all evaluation procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The study was conducted using publicly available code bases.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The study provides a detailed description of the training and hyperparameter tuning procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This study reports standard deviation across five seeds set during fine-tuning to ensure robustness of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This study provides details on computational costs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This study conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed as the study is done for research purposes.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: The data is publicly available and the models do not have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of the assets used to conduct this study are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code bases used in this study are publicly available and well documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This study does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This study does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this study does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.