A foundation model for electrodermal activity data

Anonymous Authors[1]

Abstract

Foundation models have recently extended beyond natural language and vision to time-series domains, including physiological signals. However, progress in electrodermal activity (EDA) modeling is hindered by the absence of large-scale, curated, and openly accessible datasets. EDA reflects sympathetic nervous system activity and is widely used to infer cognitive load, stress, and engagement. Yet very few wearable devices provide continuous, unobtrusive sensing, and the only large-scale archive to date is proprietary. To address this gap, we compile EDAMAME, a collection of EDA traces from 24 public datasets, comprising more than 25,000 hours from 634 users. Using this resource, we train UME, the first dedicated foundation model for EDA. In eight out of ten scenarios, UME outperforms baselines and matches generalist time-series foundation models while using 20× fewer computational resources. Our findings, however, also highlight the intrinsic challenges of EDA modeling, motivating further research to unlock its full potential.

1. Introduction

Thanks to their ability to learn general patterns from broad, diverse data and to adapt them to a wide range of downstream tasks, foundation models have attracted attention beyond their traditional applications in natural language processing and computer vision. In particular, generalist foundation models for *time series*, like Mantis (Feofanov et al., 2025) or Chronos (Ansari et al., 2024; 2025), have been proposed recently. Trained on cross-domain dataset collections, these models achieve remarkable performance on a diverse range of downstream tasks. Emerging foundation models for *physiological time series* data – e.g., for photoplethysmogram (PPG) (Abbaspourazad et al., 2024; Pillai et al., 2025; Saha et al., 2025; Ding et al., 2024; Luo et al., 2025), electrocardiogram (ECG) (Abbaspourazad et al., 2024; McKeen et al., 2025; Li et al., 2025), and electroencephalogram (EEG) data (Sukhbaatar et al., 2025) – show compelling performance, too. Yet their training remains constrained by the absence of large-scale, curated, and open datasets for physiological signals. As highlighted by Abbaspourazad et al. (2024) with respect to time series data in the medical domain, "*datasets are usually small in comparison to other domains, which is an obstacle for developing neural network models for biosignals*".

This paucity of data is particularly acute for electrodermal activity (EDA) signals. EDA refers to changes in the skin's electrical conductance caused by variations in sweat gland activity, which is itself regulated by the sympathetic branch of the autonomic nervous system. Alongside more commonly used physiological signals, EDA has numerous applications in personal informatics systems. Specifically, because EDA increases with physiological arousal, it is commonly used to assess cognitive load (Kahneman et al., 1969; Romine et al., 2020), stress (Gashi et al., 2020; Pinge et al., 2024), and engagement (Gao et al., 2020; Gashi et al., 2019; Bustos-Lopez et al., 2022). Yet EDA sensors are not (yet) commonplace on wearable devices. The only large-scale archive of EDA data is a proprietary dataset collected using the Fitbit Sense 2 (McDuff et al., 2024). This lack of (open) data has, to date, hampered the development of foundation models for EDA.

To cope with this challenge, we first constructed a large-scale archive of EDA data that integrates records from 24 different publicly available datasets and a total of more than 25'000 hours of data of 634 different users. We curated this collection of datasets, to which we refer to as EDAMAME (**EDA** **M**ulti-dataset **A**rchive for **M**odel training and **E**valuation), to train a foundation model for EDA data. The model, called UME (fo**U**ndation **M**odel for **E**lectrodermal activity data), has been trained on approximately 275 million of 60-second windows of EDA data and it is, to best of our knowledge, the first EDA-specific foundation model reported in the literature. We evaluated UME on several downstream tasks and found that it surpasses baseline models trained on both generic and EDA-specific handcrafted features in 8 out of 10 tests and matches the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

performance of generalist time series foundation models, while requiring at least 20× less computational resources. Our results, however, also highlight the intrinsic difficulty of working with EDA signals: balanced accuracy scores rarely exceed 0.7 and exhibit substantial variability. Further research is therefore needed to fully harness the potential of EDA for both unimodal and multimodal ubiquitous sensing.

We open source the code used for this work and the UME's weights (see Appendix A.5 for information about the code links). We will make the EDAMAME dataset available to other researchers upon publication of this work.

2. Background & related work

Foundation models are trained using large-scale corpora (Bommasani et al., 2022). LLMs have been enabled, in part, by the availability of text-based datasets. However, the collection large amount of physiological data requires significantly more resources and time, which has to-date limited development.

Researchers have trained foundation models for physiological data using large-scale clinical corpora, which have been made readily available by research efforts in recent years (Lee & Akamatsu, 2025). Indeed, a majority of open source foundation models have been trained on physiological signals commonly found in clinical setups: ECG, EEG or PPG. We refer to Table A.1 for an overview of existing methods.

Wearable devices enable the possibility to collect large-scale datasets from real-world users. In this context, Abbaspourazad et al. (2024) were the first to introduce foundation models trained on wearable PPG and ECG data. The authors, similar to other recent work on wearable foundation models (Saha et al., 2025; Pillai et al., 2025; Ding et al., 2024; McKeen et al., 2025; Li et al., 2025), relied on non-public datasets, collected by wearable manufacturers themselves. Only limited work, e.g., (Saha et al., 2025), exists on training foundation models for wearable data through publicly available datasets. Moreover, the limited used of EDA sensor on commercial devices means that even private datasets do not contain EDA.

We refer to Appendix A.1 for a detailed description of existing related work.

Overall, existing work suggests that foundation models for physiological data can be trained and used on a diverse set of downstream tasks when using PPG, ECG or multi-modal approaches. However, a large part of publicly available physiological foundation models rely on clinical data, and there is limited evidence that features from these models can be used with data from wearable devices and real-life settings. Recent work on multi-modal (Luo et al., 2025) and PPG (Saha et al., 2025) foundation models shows that, even without large-scale proprietary datasets, researchers are able to train open source foundation models. We create a collection of datasets, EDAMAME and, with it, we train an open source foundation model for EDA data, UME.

3. EDAMAME: a collection of electrodermal activity datasets

We address the availability of large scale datasets containing wearable EDA data through EDAMAME (**EDA** **M**ulti-dataset **A**rchive for **M**odel training and **E**valuation). EDAMAME is a collection of existing, smaller scale, datasets prepared and pre-processed in a unified manner. The goal of EDAMAME is to enable researchers to train foundation models for wearable EDA data.

We select datasets for our collection using a set of seven criteria, whose purpose is to ensure that EDAMAME contains *raw* EDA data from a diverse range of settings. In Appendix A.2.1 we detail the selection process.

We identify a potential 37 datasets. Through our selection criteria, we select a total of **24** datasets for EDAMAME, all containing EDA data Empatica E4 devices[1]. We report in Table A.2 an overview of the 24 datasets. In total, EDAMAME contains approximately 25'000 hours of EDA data from 634 users. The size of EDAMAME is in line with collections of datasets used by Saha et al. (2025) and Luo et al. (2025) to train open source foundation models for wearable physiological data. We also notice how, as reported in Appendix C.3, a smaller dataset size would not have allowed to pre-train a foundation model for EDA that would outperform existing methods.

4. UME: open source foundation model for EDA data

UME (A foUndation Model for Electrodermal activity data) is the first open source foundation model trained on wearable EDA data. With UME, our objective is to obtain a model whose internal representations encode key characteristics of the input EDA signal, such that these representations can be extracted and used as substitutes for domain-informed features in downstream tasks. We define UME following similar work on foundation models for physiological data, e.g., (Abbaspourazad et al., 2024; Pillai et al., 2025; Saha et al., 2025): we rely on self-supervised learning with a contrastive learning objective.

We show in Figure 1 an overview of the pipeline used to train and evaluate UME.

[1] https://www.empatica.com/research/e4/

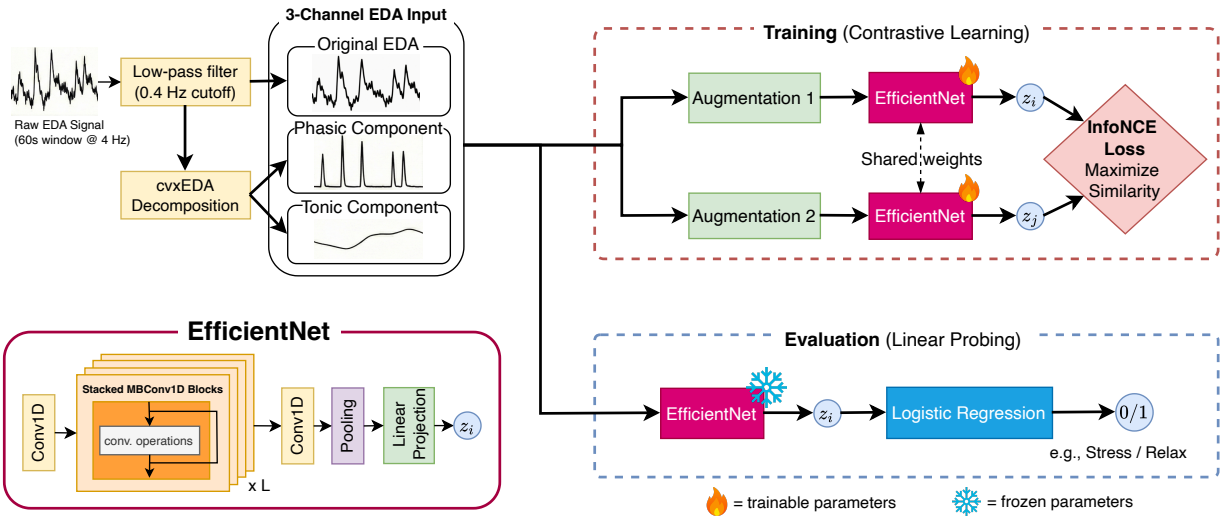


Figure 1. Overview of the train and downstream evaluation process we used for the UME foundation model. For reference, we report a sketch of the EfficientNet architecture, which we use as backbone for UME.

Data pre-processing We train UME using a subset of EDAMAME, which we call EDAMAME-train. We use the rest of the dataset, which we call EDAMAME-test, for downstream evaluation. We select 17 datasets for EDAMAME-train and 7 datasets for EDAMAME-test. With this split, the model is tested on unseen datasets.

We apply a common pre-processing pipeline to the EDA data, as in related work (Alchieri et al., 2024; Di Lascio et al., 2019; Hossain et al., 2022). We decompose the EDA into its phasic and tonic components. We use both of these components, as well as the original EDA signal, as input for our experiments. We segment the EDA data into into 60-second window with 0.25s overlap (maximum overlap), as Matton et al. (2023) and Schmidt et al. (2018). We obtain a train set consisting of approximately *275 million* windows of EDA data. For downstream evaluation, we use the same window length but non-overlapping. More details are provided in Appendix A.3.1.

Model details We chose an EfficientNet backbone for UME, similarly to related work (Abbaspourazad et al., 2024; Thapa et al., 2024); this method is computationally more efficient than traditional CNNs and, at the same time, underscores the possible use on real-world edge devices. We pretrain the model using contrastive learning, with positive pairs given by augmenting the same input sample twice. We use the EDA-specific set of data augmentation from Matton et al. (2023). We use the InfoNCE loss (van den Oord et al., 2019) as training objective. Further details are provided in Appendix A.3.2. In Appendix C we also report experiments for other backbones, model sizes and pretrain dataset sizes.

5. Evaluation of UME

We evaluate UME on the selected downstream tasks from EDAMAME-test using *linear probing*. We freeze the model weights, use it as feature extractor and then train a linear classifier. The evaluation reflects similar setups from the literature (Abbaspourazad et al., 2024; Pillai et al., 2025). We rely on two cross validation methods: Leave-One-Participant-Out (LOPO) and Time-Aware (TA) cross-validation (Alchieri et al., 2025b). We compare the features set computed from UME to various baseline features, including generic handcrafted features, a set of EDA-specific features, and features computed from generalist time series foundation models.

Finally, we also evaluate the computation complexity of the UME foundation model in extracting features, compared to the other baseline methods.

Comparison with generic handcrafted features In Figure 2 we present the results, in terms of balanced accuracy, comparing the performance of features from UME to the generic handcrafted feature set. The results show that the features from our foundation model outperform the generic handcrafted features in 9 out of 10 tasks. The improvement holds true regardless of the scenario and the task selected. We conclude from these findings that the UME foundation model learns features useful for EDA data in performing downstream predictions.

Comparison with other baseline feature sets We compare features from UME with more complex, EDA-specific handcrafted features, finding that in 8 out of 9 scenarios

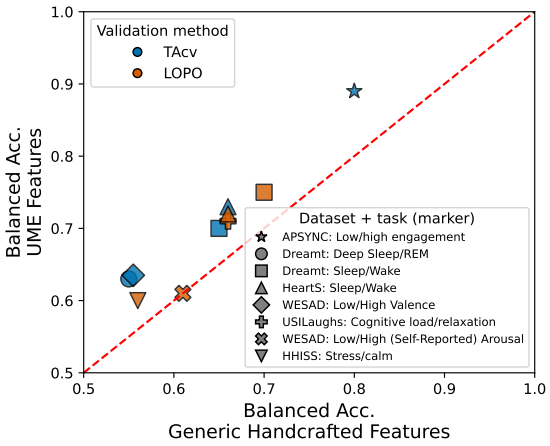


Figure 2. Pairplot with comparison in downstream tasks when using features from UME or the generic handcrafted feature set.

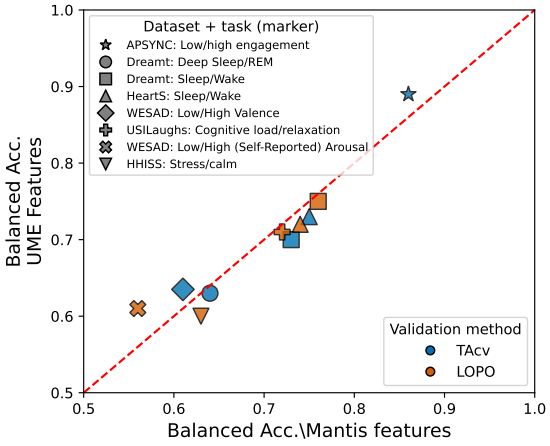


Figure 3. Pairplot with comparison in downstream tasks when using features from UME or Mantis.

our method achieves higher balanced accuracy. We also compare with generalist foundation models: Mantis (Feofanov et al., 2025), MOMENT (Goswami et al., 2024) and Chronos (Ansari et al., 2024). We find that UME outperforms, in a majority of scenarios, MOMENT and Chronos. On the other hand, Mantis, which is trained specifically for classification-based tasks, has performance similar to UME. We show the comparison with Mantis in Figure 3.

We report these results, and provide a detailed analysis, in Appendix A.4.3. Overall, we find that UME, trained on the EDAMAME collection of datasets, captures the time-series dynamics of EDA data and can be used on relevant downstream tasks.

Computational complexity We notice that the parameter size of UME is significantly smaller than the other foundation models we compare with, i.e., Mantis, MOMENT and Chronos. In terms of GFLOPs, we find that UME is $20\times$ less expensive than Mantis, the smallest generalist model, and $1500\times$ than MOMENT, the largest. Overall, UME achieves performance that is on par or higher than larger generalist models, at a fraction of the size. A detailed description of these findings is provided in Appendix A.4.4.

6. Discussion & future work

We observe that features computed using UME, generalist time-series foundation models, and EDA-specific handcrafted features lead to comparable downstream performance. We believe that the reported results are close to the intrinsic upper limit reachable for the considered downstream tasks, due to the arguably limited information in the input signal that is predictive of the label. Potential additional gains can however be obtained by coping with the significant label noise and signal noise of EDA data. Wearable EDA is noisy, strongly affected by motion/contact artifacts and low signal-to-noise ratio, in particular when collected in the wild (Gashi et al., 2020). Additionally, high-level constructs (e.g., stress, engagement, valence) might only be weakly identifiable from EDA alone without additional context. We believe that this represents an important direction for future work, e.g., by quantifying the impact of signal quality filtering, label reliability, and more. Moreover, we speculate that our results showcase how specializing foundation models for EDA data limits their size and, consequently, prospect their usage on wearable devices.

7. Conclusions

In conclusion, in this work we present EDAMAME, a large scale collection of datasets containing wearable EDA data. EDAMAME, which is composed of 24 datasets, contains approximately 25'000 hours of EDA data from about 630 users. With its diversity and variability, it enables the training of foundation models for EDA data, as we show with the training of UME, the first foundation model for wearable EDA.

Through a comprehensive set of experiments, we show that the latent representations obtained using UME outperform a set of generic handcrafted features when using linear probing in 9 out of 10 downstream tasks. We also find that UME's features achieve performance on-par with specialized EDA features and large-scale generalist foundation models, in the same set of downstream tasks. However, compared to generalist models like Chronos (Ansari et al., 2024) or Mantis (Feofanov et al., 2025), UME is computationally lighter, e.g., $20\times$ less demanding than the smallest generalist foundation model tested, enabling its possible usage on real-world wearable devices.

References

Abbaspourazad, S., Elachqar, O., Miller, A. C., Emrani, S., Nallasamy, U., and Shapiro, I. Large-scale Training of Foundation Models for Wearable Biosignals. In *International Conference on Learning Representations*, Vienna, Austria, 2024. International Conference on Learning Representations (ICLR).

Abdalazim, N., Larraza, J. A. A., Alchieri, L., Alecci, L., Santini, S., and Gashi, S. Heart Rate During Sleep Measured Using Finger-, Wrist- and Chest-Worn Devices: A Comparison Study. In Tsanas, A. and Triantafyllidis, A. (eds.), *Pervasive Computing Technologies for Healthcare*, pp. 18–32, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34586-9. doi: 10.1007/978-3-031-34586-9_2.

Abdalazim, N., Alchieri, L., Alecci, L., Barbiero, P., and Santini, S. The Impact of Domain Shift on Predicting Perceived Sleep Quality from Wearables. *Sensors*, 25 (13):4012, January 2025. ISSN 1424-8220. doi: 10.3390/s25134012.

Alchieri, L., Abdalazim, N., Alecci, L., Gashi, S., Gjoreski, M., and Santini, S. Lateralization Effects in Electrodermal Activity Data Collected Using Wearable Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1):2:1–2:30, March 2024. doi: 10.1145/3643541.

Alchieri, L., Candian, L., Abdalazim, N., Alecci, L., De Felice, G., and Santini, S. Exploring Generalist Foundation Models for Time Series of Electrodermal Activity Data. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp Companion '25, pp. 885–892, New York, NY, USA, December 2025a. Association for Computing Machinery. ISBN 979-8-4007-1477-1. doi: 10.1145/3714394.3756186.

Alchieri, L., Scocco, V., Abdalazim, N., Alecci, L., and Santini, S. Improving Human Behavior Recognition Using Bilateral Electrodermal Activity Data Collected From Wearable Devices. In *2025 International Conference on Activity and Behavior Computing (ABC)*, pp. 1–10, Al Ain, United Arab Emirates, April 2025b. IEEE. doi: 10.1109/ABC64332.2025.11118532.

Almadhor, A., Sampedro, G. A., Abisado, M., Abbas, S., Kim, Y.-J., Khan, M. A., Baili, J., and Cha, J.-H. Wrist-Based Electrodermal Activity Monitoring for Stress Detection Using Federated Learning. *Sensors*, 23(8):3984, January 2023. ISSN 1424-8220. doi: 10.3390/s23083984.

Ansari, A. F., Stella, L., Turkmen, A. C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, B. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, May 2024. ISSN 2835-8856.

Ansari, A. F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S. S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D. C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P. M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., and Bohlke-Schneider, M. Chronos-2: From Univariate to Universal Forecasting, October 2025.

Apertus, P., Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., Hakimi, I., Giraldo, J. G., Ismayilzada, M., Foroutan, N., Moalla, S., Chen, T., Sabolčec, V., Xu, Y., Aerni, M., AlKhamissi, B., Mariñas, I. A., Amani, M. H., Ansaripour, M., Badanin, I., Benoit, H., Boros, E., Browning, N., Bösch, F., Böther, M., Canova, N., Challier, C., Charmillot, C., Coles, J., Deriu, J., Devos, A., Drescher, L., Dzenhaliou, D., Ehrmann, M., Fan, D., Fan, S., Gao, S., Gila, M., Grandury, M., Hashemi, D., Hoyle, A., Jiang, J., Klein, M., Kucharavy, A., Kucherenko, A., Lübeck, F., Machacek, R., Manitaras, T., Marfurt, A., Matoba, K., Matrenok, S., Mendonça, H., Mohamed, F. R., Montariol, S., Mouchel, L., Najem-Meyer, S., Ni, J., Oliva, G., Pagliardini, M., Palme, E., Panferov, A., Paoletti, L., Passerini, M., Pavlov, I., Poiroux, A., Ponkshe, K., Ranchin, N., Rando, J., Sauser, M., Saydaliev, J., Sayfiddinov, M. A., Schneider, M., Schuppli, S., Scialanga, M., Semenov, A., Shridhar, K., Singhal, R., Sotnikova, A., Sternfeld, A., Tarun, A. K., Teiletche, P., Vamvas, J., Yao, X., Zhao, H., Ilic, A., Klimovic, A., Krause, A., Gulcehre, C., Rosenthal, D., Ash, E., Tramèr, F., VandeVondele, J., Veraldi, L., Rajman, M., Schulthess, T., Hoefler, T., Bosselut, A., Jaggi, M., and Schlag, I. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments, December 2025.

Bent, B., Cho, P. J., Henriquez, M., Wittmann, A., Thacker, C., Feinglos, M., Crowley, M. J., and Dunn, J. P. Engineering digital biomarkers of interstitial glucose from noninvasive smartwatches. *npj Digital Medicine*, 4 (1):89, June 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00465-w.

Beten, A., Lococco, L., Baig, A., and Karunarathna, T. Systematic Analysis of Distribution Shifts in Cross-Subject Glucose Prediction Using Wearable Physiological Data. *Engineering Proceedings*, 118(1):88, 2025. ISSN 2673-4591. doi: 10.3390/ECSA-12-26583.

Bizzego, A., Gabrieli, G., Furlanello, C., and Esposito, G.

Comparison of Wearable and Clinical Devices for Acquisition of Peripheral Nervous System Signals. *Sensors (Basel, Switzerland)*, 20(23):6778, November 2020. ISSN 1424-8220. doi: 10.3390/s20236778.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the Opportunities and Risks of Foundation Models, July 2022.

Boucsein, W. *Electrodermal Activity*. Electrodermal Activity, 2nd Ed. Springer Science + Business Media, New York, NY, US, 2012. ISBN 978-1-4614-1125-3 978-1-4614-1126-0. doi: 10.1007/978-1-4614-1126-0.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124, Istanbul, Turkey, August 2010. IEEE. doi: 10.1109/ICPR.2010.764.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, Online, 2020. Curran Associates, Inc.

Bukharin, A., Li, S., Wang, Z., Yang, J., Yin, B., Li, X., Zhang, C., Zhao, T., and Jiang, H. Data Diversity Matters for Robust Instruction Tuning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3411–3425, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.195.

Bustos-Lopez, M., Cruz-Ramirez, N., Guerra-Hernandez, A., Sánchez-Morales, L. N., Cruz-Ramos, N. A., and Alor-Hernandez, G. Wearables for engagement detection in learning environments: A review. *Biosensors*, 12(7): 509, 2022.

Campanella, S., Altaleb, A., Belli, A., Pierleoni, P., and Palma, L. A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. *Sensors*, 23(7):3565, January 2023. ISSN 1424-8220. doi: 10.3390/s23073565.

Casanovas Ortega, M., Bruno, E., and Richardson, M. P. Electrodermal activity response during seizures: A systematic review and meta-analysis. *Epilepsy & Behavior: E&B*, 134:108864, September 2022. ISSN 1525-5069. doi: 10.1016/j.yebeh.2022.108864.

Chen, J., Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., and Chan, S.-H. G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12021–12031, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01157.

Chen, Z., Wang, S., Xiao, T., Wang, Y., Chen, S., Cai, X., He, J., and Wang, J. Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23881–23899, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1163.

Chicco, D. and Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21 (1):6, January 2020. ISSN 1471-2164. doi: 10.1186/s12864-019-6413-7.

Coffman, D. L., Cai, X., Li, R., and Leonard, N. R. Challenges and Opportunities in Collecting and Modeling Ambulatory Electrodermal Activity Data. *JMIR biomedical engineering*, 5(1):e17106, 2020. ISSN 2561-3278. doi: 10.2196/17106.

Dehghani, A., Sarbishei, O., Glatard, T., and Shihab, E. A Quantitative Comparison of Overlapping and Non-Overlapping Sliding Windows for Human Activity Recognition Using Inertial Sensors. *Sensors*, 19(22):5026, January 2019. ISSN 1424-8220. doi: 10.3390/s19225026.

Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006. ISSN 1532-4435.

Derdiyok, S., Akbulut, F. P., and Catal, C. Neurophysiological and biosignal data for investigating occupational mental fatigue: MEFAR dataset. *Data in Brief*, 52:109896, February 2024. ISSN 2352-3409. doi: 10.1016/j.dib.2023.109896.

Di Lascio, E., Gashi, S., and Santini, S. Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3):103:1–103:21, September 2018. doi: 10.1145/3264913.

Di Lascio, E., Gashi, S., and Santini, S. Laughter Recognition Using Non-invasive Wearable Devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth'19, pp. 262–271, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 978-1-4503-6126-2. doi: 10.1145/3329189.3329216.

Di Lascio, E., Gashi, S., Debus, M. E., and Santini, S. Automatic Recognition of Flow During Work Activities Using Context and Physiological Signals. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, Nara, Japan, September 2021. IEEE. doi: 10.1109/ACII52823.2021.9597434.

Ding, C., Guo, Z., Chen, Z., Lee, R. J., Rudin, C., and Hu, X. SiamQuality: A ConvNet-based foundation model for photoplethysmography signals. *Physiological Measurement*, 45(8):085004, August 2024. ISSN 0967-3334. doi: 10.1088/1361-6579/ad6747.

Dissanayake, V., Seneviratne, S., Rana, R., Wen, E., Kaluarachchi, T., and Nanayakkara, S. SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning. *IEEE Access*, 10:18105–18120, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3149509.

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. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, October 2020. International Conference on Learning Representations (ICLR).

Feofanov, V., Wen, S., Alonso, M., Ilbert, R., Guo, H., Tiomoko, M., Pan, L., Zhang, J., and Redko, I. Mantis: Lightweight Calibrated Foundation Model for User-Friendly Time Series Classification, February 2025.

Gao, N., Shao, W., Rahaman, M. S., and Salim, F. D. N-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3):79:1–79:26, September 2020. doi: 10.1145/3411813.

Garcia-Moreno, F. M. and Badenes-Sastre, M. EmoPulse Moments E4 Dataset (EPM-E4): An Exhaustive Collection of Emotion-Related Data from Empatica E4 Wearable, September 2020.

Gashi, S., Di Lascio, E., and Santini, S. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(1):13:1–13:19, March 2019. doi: 10.1145/3314400.

Gashi, S., Di Lascio, E., Stancu, B., Swain, V. D., Mishra, V., Gjoreski, M., and Santini, S. Detection of Artifacts in Ambulatory Electrodermal Activity Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2):44:1–44:31, June 2020. doi: 10.1145/3397316.

Gashi, S., Alecci, L., Lascio, E. D., Debus, M. E., Gasparini, F., and Santini, S. The Role of Model Personalization for Sleep Stage and Sleep Quality Recognition Using Wearables. *IEEE Pervasive Computing*, 21(2):69–77, April 2022a. ISSN 1558-2590. doi: 10.1109/MPRV.2022.3164334.

Gashi, S., Min, C., Montanari, A., Santini, S., and Kawsar, F. A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices. *Scientific Data*, 9(1):537, September 2022b. ISSN 2052-4463. doi: 10.1038/s41597-022-01643-5.

Gjoreski, M., Kolenik, T., Knez, T., Luštrek, M., Gams, M., Gjoreski, H., and Pejović, V. Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits. *Applied Sciences*, 10(11):3843, January 2020. ISSN 2076-3417. doi: 10.3390/app10113843.

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., and Dubrawski, A. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 16115–16152, Vienna, Austria, July 2024. JMLR.org.

Grandini, M., Bagli, E., and Visani, G. Metrics for Multi-Class Classification: An Overview, August 2020.

Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., and Citi, L. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, April 2016. ISSN 1558-2531. doi: 10.1109/TBME.2015.2474131.

Hossain, M.-B., Kong, Y., Posada-Quintero, H. F., and Chon, K. H. Comparison of Electrodermal Activity from Multiple Body Locations Based on Standard EDA Indices' Quality and Robustness against Motion Artifact. *Sensors*, 22(9):3177, January 2022. ISSN 1424-8220. doi: 10.3390/s22093177.

Hossain, M. B., Kong, Y., Posada-Quintero, H., and Chon, K. *Electrodermal Activity: Applications and Challenges*, pp. 475–496. Cambridge University Press, November 2024. ISBN 978-1-316-51855-7. doi: 10.1017/9781009000796.022.

Hosseini, S., Gottumukkala, R., Katragadda, S., Bhupatiraju, R. T., Ashkar, Z., Borst, C. W., and Cochran, K. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data*, 9(1):255, June 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01361-y.

Iqbal, T., Simpkin, A. J., Roshan, D., Glynn, N., Killilea, J., Walsh, J., Molloy, G., Ganly, S., Ryman, H., Coen, E., Elahi, A., Wijns, W., and Shahzad, A. Stress Monitoring Using Wearable Sensors: A Pilot Study and Stress-Predict Dataset. *Sensors*, 22(21):8135, January 2022. ISSN 1424-8220. doi: 10.3390/s22218135.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x.

Kahneman, D., Tursky, B., Shapiro, D., and Crider, A. Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of experimental psychology*, 79(1p1):164, 1969.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization, 2017.

Koscova, Z., Moura Junior, V., Reyna, M., Hong, S., Gupta, A., Ghanta, M., Sameni, R., Aguirre, A., Li, Q., Zafar, S., Clifford, G., and Westover, M. B. Harvard-Emory ECG Database, 2025.

Laporte, M., Gjoreski, M., and Langheinrich, M. LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(3):106:1–106:41, September 2023. doi: 10.1145/3610892.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3031549.

Lee, H.-C., Park, Y., Yoon, S. B., Yang, S. M., Park, D., and Jung, C.-W. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, June 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01411-5.

Lee, S. and Akamatsu, K. Foundation Models for Physiological Signals: Opportunities and Challenges, August 2025.

Li, J., Aguirre, A. D., Moura, V., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Westover, M. B., and Hong, S. An Electrocardiogram Foundation Model Built on over 10 Million Recordings. *Nejm Ai*, 2(7):AIoa2401033, July 2025. ISSN 2836-9386. doi: 10.1056/aioa2401033.

Liu, Y., Palacio, M.-I., Bikki, T., Toledo, C., Ouyang, Y., Li, Z., Wang, Z., Toledo, F., Zeng, H., and Herrero, M.-T. Machine Learning, Physiological Signals, and Emotional Stress/Anxiety: Pitfalls and Challenges. *Applied Sciences*, 15(21):11777, January 2025. ISSN 2076-3417. doi: 10.3390/app152111777.

Lukan, J., Gjoreski, M., Mauersberger, H., Hoppe, A., Hess, U., and Luštrek, M. Analysing Physiology of Interpersonal Conflicts Using a Wrist Device. In Kameas, A. and Stathis, K. (eds.), *Ambient Intelligence*, pp. 162–167, Cham, 2018. Springer International Publishing. ISBN 978-3-030-03062-9. doi: 10.1007/978-3-030-03062-9_13.

Luo, Y., Chen, Y., Salekin, A., and Rahman, T. Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals, May 2025.

Lutin, E., Hashimoto, R., De Raedt, W., and Van Hoof, C. Feature Extraction for Stress Detection in Electrodermal Activity:. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 177–185, Online Streaming, — Select a Country —, 2021. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-490-9. doi: 10.5220/0010244601770185.

Matton, K., Lewis, R., Guttag, J., and Picard, R. Contrastive Learning of Electrodermal Activity Representations for Stress Detection. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 410–426, Cambridge, MA, USA, June 2023. PMLR.

McDuff, D., Thomson, S., Abdel-Ghaffar, S., Galatzer-Levy, I. R., Poh, M.-Z., Sunshine, J., Barakat, A., Heneghan, C., and Sunden, L. What Does Large-scale Electrodermal Sensing Reveal?, February 2024.

McKeen, K., Masood, S., Toma, A., Rubin, B., and Wang, B. ECG-FM: An open electrocardiogram foundation model. *JAMIA open*, 8(5):ooaf122, October 2025. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooaf122.

Mishra, V., Sen, S., Chen, G., Hao, T., Rogers, J., Chen, C.-H., and Kotz, D. Evaluating the Reproducibility of Physiological Stress Detection Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4):147:1–147:29, December 2020. doi: 10.1145/3432220.

Narayanswamy, G., Liu, X., Ayush, K., Yang, Y., Xu, X., Liao, S., Garrison, J., Tailor, S., Sunshine, J., Liu, Y., Althoff, T., Narayanan, S., Kohli, P., Zhan, J., Malhotra, M., Patel, S., Abdel-Ghaffar, S., and McDuff, D. Scaling Wearable Foundation Models, October 2024.

Neupane, S., Saha, M., Ali, N., Hnat, T., Samiei, S. A., Nandugudi, A., Almeida, D. M., and Kumar, S. Momentary Stressor Logging and Reflective Visualizations: Implications for Stress Management with Wearables. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–19, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3642662.

Owusu-Adjei, M., Hayfron-Acquah, J. B., Frimpong, T., and Abdul-Salaam, G. Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, 2 (11):e0000290, November 2023. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000290.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, pp. 8026–8037. Curran Associates Inc., Red Hook, NY, USA, 2019.

Pillai, A., Spathis, D., Kawsar, F., and Malekzadeh, M. PaPaGei: Open Foundation Models for Optical Physiological Signals, February 2025.

Pinge, A., Gad, V., Jaisighani, D., Ghosh, S., and Sen, S. Detection and monitoring of stress using wearables: a systematic review. *Frontiers in Computer Science*, 6: 1478851, 2024.

Poh, M.-Z., Loddenkemper, T., Reinsberger, C., Swenson, N. C., Goyal, S., Sabtala, M. C., Madsen, J. R., and Picard, R. W. Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor. *Epilepsia*, 53(5):e93–97, May 2012. ISSN 1528-1167. doi: 10.1111/j.1528-1167.2012.03444.x.

Rafiul Amin, Md., Wickramasuriya, D. S., and Faghih, R. T. A Wearable Exam Stress Dataset for Predicting Grades using Physiological Signals. In *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pp. 30–36, Houston, TX, USA, March 2022. IEEE. doi: 10.1109/HI-POCT54491.2022.9744065.

Rehman, S. U., Ali, A., Khan, A. M., and Okpala, C. Human Activity Recognition: A Comparative Study of Validation Methods and Impact of Feature Extraction in Wearable Sensors. *Algorithms*, 17(12):556, December 2024. ISSN 1999-4893. doi: 10.3390/a17120556.

Reiss, A., Indlekofer, I., Schmidt, P., and Van Laerhoven, K. Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors*, 19(14):3079, January 2019. ISSN 1424-8220. doi: 10.3390/s19143079.

Reyna, M. A., Sadr, N., Alday, E. A. P., Gu, A., Shah, A. J., Robichaux, C., Rad, A. B., Elola, A., Seyedi, S., Ansari, S., Ghanbari, H., Li, Q., Sharma, A., and Clifford, G. D. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. In *2021 Computing in Cardiology (CinC)*, volume 48, pp. 1–4, Brno, Czech Republic, September 2021. IEEE. doi: 10.23919/CinC53138.2021.9662687.

Romine, W. L., Schroeder, N. L., Graft, J., Yang, F., Sadeghi, R., Zabihimayvan, M., Kadariya, D., and Banerjee, T. Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: Development of a cognitive load tracker for both personal and classroom use. *Sensors*, 20 (17):4833, 2020.

Saeed, A., Salim, F. D., Ozcelebi, T., and Lukkien, J. Federated Self-Supervised Learning of Multisensor Representations for Embedded Intelligence. *IEEE Internet of Things Journal*, 8(2):1030–1040, January 2021. ISSN 2327-4662. doi: 10.1109/JIOT.2020.3009358.

Saha, M., Xu, M. A., Mao, W., Neupane, S., Rehg, J. M., and Kumar, S. Pulse-PPG: An Open-Source Field-Trained PPG Foundation Model for Wearable Applications across Lab and Field Settings. *Proc. ACM Interact.*

Mob. Wearable Ubiquitous Technol., 9(3):126:1–126:35, September 2025. doi: 10.1145/3749494.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pp. 400–408, New York, NY, USA, October 2018. Association for Computing Machinery. ISBN 978-1-4503-5692-3. doi: 10.1145/3242969.3242985.

Shuai, Z., Xu, Z., Yang, D., Wang, W., and Yang, Y. OSF: On Pre-training and Scaling of Sleep Foundation Models, February 2026. URL http://arxiv.org/abs/2603.00190. arXiv:2603.00190 [cs].

Slade, C., Sun, Y., Chao, W. C., Chen, C.-C., Benzo, R. M., and Washington, P. Current challenges and opportunities in active and passive data collection for mobile health sensing: A scoping review. *JAMIA open*, 8(4):ooaf025, August 2025. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooaf025.

Sukhbaatar, J., Imamura, S., Inoue, I., Murakami, S., Hassan, K. M., Han, S., Chanpornpakdi, I., and Tanaka, T. SingLEM: Single-Channel Large EEG Model, September 2025.

Svoren, H., Thambawita, V., Halvorsen, P., Jakobsen, P., Garcia-Ceja, E., Noori, F. M., Hammer, H. L., Lux, M., Riegler, M. A., and Hicks, S. A. Toadstool: A dataset for training emotional intelligent machines playing Super Mario Bros. In *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20, pp. 309–314, New York, NY, USA, May 2020. Association for Computing Machinery. ISBN 978-1-4503-6845-2. doi: 10.1145/3339825.3394939.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114, Long Beach, CA, USA, 2019. PMLR, PMLR.

Tello, A., Degeler, V., and Lazovik, A. Too good to be true: Accuracy overestimation in (re) current practices for human activity recognition. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*, pp. 511–517, Biarritz, France, 2024. IEEE.

Thapa, R., He, B., Kjaer, M. R., Moore, H., Ganjoo, G., Mignot, E., and Zou, J. SleepFM: Multi-modal Representation Learning for Sleep Across Brain Activity, ECG and Respiratory Signals, May 2024. URL http://arxiv.org/abs/2405.17766. arXiv:2405.17766 [cs].

Truslow, J., Spillane, A., Lin, H., Cyr, K., Ullal, A., Arnold, E., Huang, R., Rhodes, L., Block, J., Stark, J., Kretlow, J., Beatty, A. L., Werdich, A., Bankar, D., Bianchi, M., Shapiro, I., Villalpando, J., Ravindran, S., Mance, I., Phillips, A., Earl, J., Deo, R. C., Desai, S. A., and MacRae, C. A. Understanding activity and physiology at scale: The Apple Heart & Movement Study. *npj Digital Medicine*, 7(1):242, September 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01187-5.

van den Oord, A., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding, January 2019.

Viana-Matesanz, M. and Sánchez-Ávila, C. Adaptive Normalization and Feature Extraction for Electrodermal Activity Analysis. *Mathematics*, 12(2):202, January 2024. ISSN 2227-7390. doi: 10.3390/math12020202.

Wang, K., Yang, J., Shetty, A., and Dunn, J. DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology, 2024.

Yi, R., Cao, T., Zhou, A., Ma, X., Wang, S., and Xu, M. Boosting DNN Cold Inference on Edge Devices. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, MobiSys '23, pp. 516–529, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 979-8-4007-0110-8. doi: 10.1145/3581791.3596842.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. The National Sleep Research Resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, October 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy064.

Zhang, Y., Ayush, K., Qiao, S., Heydari, A. A., Narayanswamy, G., Xu, M. A., Metwally, A. A., Xu, S., Garrison, J., Xu, X., Althoff, T., Liu, Y., Kohli, P., Zhan, J., Malhotra, M., Patel, S., Mascolo, C., Liu, X., McDuff, D., and Yang, Y. SensorLM: Learning the Language of Wearable Sensors, June 2025.

A. Appendix

A.1. Background & related work

In this section, we provide an overview of the background and relevant related work. We discuss the relevant background associated with wearable devices and EDA data. Then, we describe existing research on foundation models for physiological data, with a focus on signals collected from wearable devices. We also present work on the use of self-supervised learning applied to EDA data, since self-supervised learning is one of the main building blocks for training foundation models.

Background Foundation models require large scale datasets to be trained (Bommasani et al., 2022). Researchers have obtained state-of-the-art results in the NLP domain thanks in part to the availability of large scale textual corpora, e.g., (Apertus et al., 2025; Brown et al., 2020). Researchers create textual corpora by, for example, crawling websites and gathering all available text, e.g., as done by the Common Crawl dataset[2]. On the other hand, collecting physiological signals requires the use of specialized equipment as well as significant human and time resources (Liu et al., 2025; Slade et al., 2025). Researchers have explored foundation models for physiological data through the use of large scale clinical corpora (Lee & Akamatsu, 2025). Only recent work, e.g., (Abbaspourazad et al., 2024), explores the use of foundation models trained from real-world wearable physiological signals. This is due to significantly higher operational difficulties in gathering high-quality data when collecting wearable data, compared to clinical studies (Bizzego et al., 2020). Additionally, EDA, even if used in the healthcare domain to detect seizures (Casanovas Ortega et al., 2022), is not as commonly collected in clinical settings as PPG or ECG (Coffman et al., 2020; Poh et al., 2012).

Foundation models for physiological data In Table A.1 we provide an overview of a selected set of existing foundation models for physiological data, their size, their availability and information about the data used to train them. Current research focuses on PPG, ECG, or multi-modal approaches. A majority of the selected foundation models relies on clinical data, given its abundance. To the best of our knowledge, Saha et al. (2025)'s work is the only one using exclusively wearable physiological data to train an open source foundation model.

Researchers have explored foundation models for physiological data using both clinical and wearable data. Abbaspourazad et al. (2024) were the first to use a large scale, private, dataset of wearable data to train foundation models on PPG and ECG data. Their results show that features from foundation models can be used to perform multiple downstream tasks, with performance on par to existing approaches. Following these results, other researchers, e.g., (Saha et al., 2025; Pillai et al., 2025; Ding et al., 2024; McKeen et al., 2025; Li et al., 2025), have trained foundation models for either PPG or ECG data. However, most of the existing approaches rely on clinical PPG and ECG signals, with limited work on real-world physiological data collected from wearable devices (Saha et al., 2025), and, especially, no work on EDA data.

Multi-modal approaches have also recently been proposed. Narayanswamy et al. (2024) trained a family of multimodal foundation models on data aggregated over 1 minute. They compute aggregated features from PPG (e.g., mean heart rate), EDA (e.g., mean EDA values), and other physiological and behavioral signals. They find that their foundation model significantly outperforms (up to 50% improvement) over baseline methods. However, their approach relies on proprietary data and the authors released neither the code nor the weights for their model. Using publicly available data, Luo et al. (2025) trained a multi-modal foundation model for physiological data, e.g., PPG, ECG, EDA, from a curated collection of datasets with about 15'000 hours of data. In their experimental setup, they show promising zero-shot downstream performance for their foundation model. However, the authors' focus was not on EDA data, which means that their collection of datasets only contained a limited amount of EDA, compared to other physiological signals. Overall, a curated EDA dataset collection has not yet been made available to the wider research community.

Self-supervised learning for EDA data Researchers have also explored the use of self-supervised learning on wearable data, including EDA. Dissanayake et al. (2022) trained a self-supervised learning model, through contrastive learning, for emotion estimation using PPG, EDA and skin temperature. Saeed et al. (2021) similarly used a multi-modal self-supervised approach in a federated learning settings. Recently, Matton et al. (2023) trained a self-supervised learning model on EDA data to perform stress classification. They used contrastive learning and data augmentation techniques to achieve state-of-the-art performance on in-distribution downstream tasks.

[2] https://commoncrawl.org
[3] Datasets available, not collection.

Table A.1. Foundation models for physiological data and their training datasets.

Model				Training Data				
Ref	**Sig**	**#Param**	**Avail**	**Dataset(s)**	**Type**	**Hrs (k)**	**Users**	**Avail**
CLINICAL								
				VitalDB (Lee et al., 2022)	Clin	∼17	∼6k	✓
(Pillai et al., 2025)	PPG	∼30–150M	Open	MIMIC-III (Johnson et al., 2016)	Clin	∼20	∼6k	✓
				MESA (Zhang et al., 2018)	Clin	∼20	∼2k	✓
(Ding et al., 2024)	PPG	∼1–8M	Code	(Ding et al., 2024)	Clin	∼300	∼29k	✓
				MIMIC-IV (Johnson et al., 2023)	Clin	∼140	∼160k	✓
(McKeen et al., 2025)	ECG	∼300M	Open					
				PhysioNet21 (Reyna et al., 2021)	Clin	∼14	N/A	✓
				UHN-ECG (McKeen et al., 2025)	Clin	∼100	∼180k	✓
(Li et al., 2025)	ECG	∼30M	Open	HEEDB (Koscova et al., 2025)	Clin	∼1000	∼2M	✓
(Sukhbaatar et al., 2025)	EEG	N/A	Open	curated	Clin	∼350	∼9	✗[3]
WEARABLE								
(Abbaspourazad et al., 2024)	PPG	∼3M	Private	AHMS (Truslow et al., 2024)	Wear	∼300	∼141k	✗
	ECG	∼3M	Private			∼30	∼106k	✗
(Narayanswamy et al., 2024)	Multi	∼1–100M	Private	(Narayanswamy et al., 2024)	Wear	∼40k	∼165k	✗
(Saha et al., 2025)	PPG	∼30M	Open	MOODS (Neupane et al., 2024)	Wear	∼40	∼120	✗
OTHER								
(Luo et al., 2025)	Multi	∼140M	Open	curated	Mixed	∼15	N/A	✗[3]
OURS								
Ours	**EDA**	∼1M	**Open**	**EDAMAME**	**Wear**	∼25	∼630	✓

A.2. EDAMAME: a collection of electrodermal activity datasets

In this section, we provide a detailed description of the creation of EDAMAME and an analysis of the data distribution.

A.2.1. Description of the collection of datasets

Our goal is to create a large-scale and diverse corpus to address the scarcity of EDA-specific collections. To this end, we first identify a set of scenarios that are relevant for EDA data through relevant literature (Hossain et al., 2024). The scenarios are: *sleep monitoring*, *stress/emotion induction*, *engagement*, *real-world stress*, *workplace analysis*, *daily living*. Then, we select and search datasets from the literature using the following criteria:

1. datasets have to contain *raw* EDA data from wearable devices;
2. the individual datasets have to be either open source or available to researchers upon signing a data sharing agreement;
3. datasets have to contain data collected during tasks or moments that influence EDA signals;
4. the collection has to contain datasets collected using different protocols. e.g., lab environment or in-the-wild collection;
5. the collection needs to contain at least one dataset from each one of the six scenarios defined;
6. in order to limit the search size, we also add a saturation criterion: once we reach the threshold of 100 users and 1000 hours for one of the six scenarios, we stop searching for additional data;
7. the total size of EDAMAME has to be more than 15'000 hours and more than 100 users, which is a size similar to that of datasets used to train existing open source foundation models for wearable data (Saha et al., 2025; Luo et al., 2025).

In order to satisfy the first criterion, we collect data only from datasets using Empatica E4 devices[4]. We choose to use only Empatica E4 data since it is one of the few research-grade wearable devices that allow to continuously collect wearable EDA.

We search datasets through open source databases containing physiological data, specifically PhysioNet[5], Zotero[6], Kaggle[7]. We also search for dataset information using databases of scientific articles, e.g., Google Scholar[8], ACM Digital Library[9]. Finally, we also search through the online databases of scientific journals that publish datasets containing physiological data, e.g., Nature Scientific Data[10], IMWUT[11]. We identify a potential 37 datasets. Through the aforementioned criteria, we select a total of **24** datasets for EDAMAME. We report in Table A.2 an overview of the 24 datasets. In total, EDAMAME contains approximately 25'000 hours of EDA data from 634 users. The size of EDAMAME is in line, as outlined by our seventh selection criterion, with collections of datasets used by Saha et al. (2025); Luo et al. (2025) to train open source foundation models for wearable physiological data.

The EDA signal in all 24 datasets is sampled at 4 Hz, since this is the default sampling rate for EDA from the Empatica E4[12]. All timestamps are converted to UTC time. Whenever timestamps are missing, we assign unix-time 0 to the start of a timeseries of EDA data.

A.2.2. Data variability in EDAMAME

In this section, we show the data variability in the EDAMAME collection of datasets. We highlight the diversity of EDA data in EDAMAME to highlight its feasibility in training foundation models for EDA data. In particular, we discuss the EDA data distribution and the distribution of users and data per user across the datasets. Figure A.1 shows six example EDA signals from different datasets, representing six distinct scenarios: *sleep monitoring*, *stress/emotion induction*, *engagement*, *real-world stress*, *workplace analysis*, *daily living*.

[4] https://www.empatica.com/research/e4/
[5] https://physionet.org/
[6] https://zenodo.org/
[7] https://www.kaggle.com
[8] https://scholar.google.com
[9] https://dl.acm.org
[10] https://www.nature.com/sdata/
[11] https://dl.acm.org/journal/imwut
[12] https://www.empatica.com/blog/decoding-wearable-sensor-signals-what-to-expect-from-your-e4-data

Table A.2. Summary of the datasets included in the training split. The collection spans diverse physiological scenarios and environments.

Dataset Name	Duration (h)	# Users	Scenario	Environment
APSync (Gashi et al., 2019)	168	27	Real-world Stress	Wild
BIG IDEAS (Bent et al., 2021)	2607	16	Daily Living	Wild
BiHeartS (Abdalazim et al., 2025)	2911	11	Sleep Monitoring	Wild
DREAMT (Wang et al., 2024)	882	100	Daily Living	Wild
Dynamics in the workplace (Lukan et al., 2018)	1813	42	Workplace Analysis	Wild
EmpaticaE4Stress (Campanella et al., 2023)	5	29	Stress/Emotion Induction	Lab
EPM-E4 (Garcia-Moreno & Badenes-Sastre, 2020)	22	47	Engagement	Wild
HeartS (Abdalazim et al., 2023)	886	5	Sleep Monitoring	Wild
HHISS (Gjoreski et al., 2020)	3166	46	Real-world Stress	Wild
LAUREATE (Laporte et al., 2023)	1406	46	Engagement	Wild
M2Sleep (Gashi et al., 2022a)	8403	16	Sleep Monitoring	Wild
MEFAR (Derdiyok et al., 2024)	27	23	Stress/Emotion Induction	Lab
Nurses' Stress (Hosseini et al., 2022)	800	18	Workplace Analysis	Wild
PPG-Dalia (Reiss et al., 2019)	38	15	Daily Living	Wild
SEED (Di Lascio et al., 2018)	340	31	Engagement	Wild
SEED-II-Lab (Di Lascio et al., 2018)	29	25	Engagement	Lab
SEED-II-Wild (Di Lascio et al., 2018)	156	6	Engagement	Wild
Stress Predict (Iqbal et al., 2022)	32	35	Stress/Emotion Induction	Lab
ToadStool (Svoren et al., 2020)	10	10	Stress/Emotion Induction	Lab
USILaughs (Di Lascio et al., 2019)	1.5	30	Stress/Emotion Induction	Lab
WEEE (Gashi et al., 2022b)	17	17	Daily Living	Lab
WESAD (Schmidt et al., 2018)	29	15	Stress/Emotion Induction	Lab
WESD (Rafiul Amin et al., 2022)	123	10	Real-world Stress	Wild
Workplace (Di Lascio et al., 2021)	830	14	Workplace Analysis	Wild
Total	**24'735**	**634**		

Data distribution Figure A.2 shows the distribution of raw EDA values across the 24 datasets in EDAMAME. We highlight how data distribution is similar across datasets that are collected in similar scenarios. Higher EDA values are associated with stronger ANS arousal activation (Boucsein, 2012). For example, datasets in the *Stress/Emotion Induction* scenario have longer tails, in their distribution, compared to others. On the other hand, datasets in the *Real-world Stress* scenario have lower tails and more EDA values close to 0 μS. These two distinct patterns highlight the challenges of real-world EDA data: a lower distribution of EDA values in the *Real-world Stress* scenario suggests that during daily life less arousal-associated moments are present. Overall, EDAMAME contains diverse data, across both scenarios and EDA values. This diversity is important when training foundation models, since it exposes the model to a wide range of EDA data.

Number of users and data per user Figure A.3 shows the number of users compared to the data density per user, for each dataset in EDAMAME. The plot shows the structural topology of the collection of datasets. Specifically, the collection contains two types of datasets: datasets that contain less users but more data per user (longitudinal depth); and datasets that contain more users, but less data per user (high population diversity). The first set of datasets in which multiple days of data for each user are present contains more intra-personal variability information, i.e., how people's data changes over time, compared to the second set. The second set, on the other hand, contains more inter-personal variability information, i.e., how data differs between different participants, compared to the first one.

EDAMAME contains data from a diverse range of scenarios and users. The collection also balances datasets with longitudinal depth and datasets with high population diversity. This diversity and variability is in line with recent literature on foundation model train data (Chen et al., 2025; Bukharin et al., 2024).

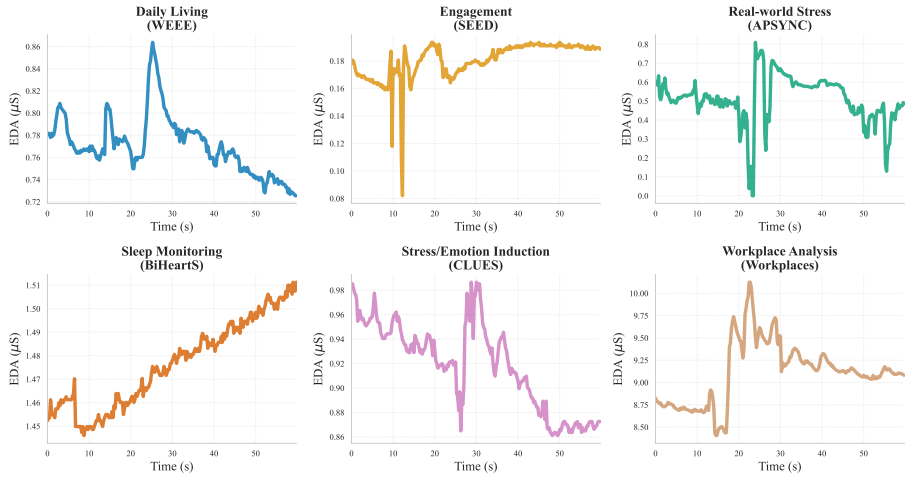


Figure A.1. Example of six EDA signals (randomly drawn). Each signal is from a different dataset, each for one of the six scenarios.

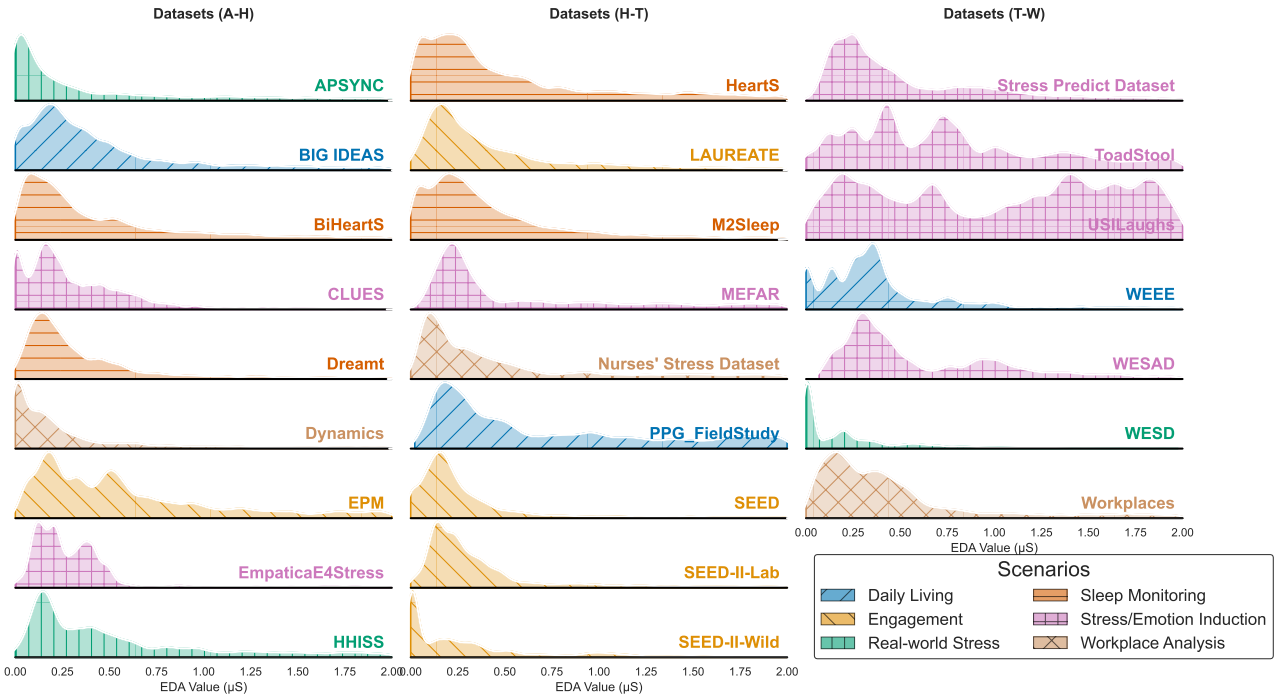


Figure A.2. Ridgeline plot with distribution of EDA values from each dataset. The colors represent the six scenarios each dataset was collected in. The plot has been truncated, for visualization's sake, in the range $0 - 2.5$ μS. The original range was $0 - 40$ μS (the E4's max theoretical value is 100 μS).

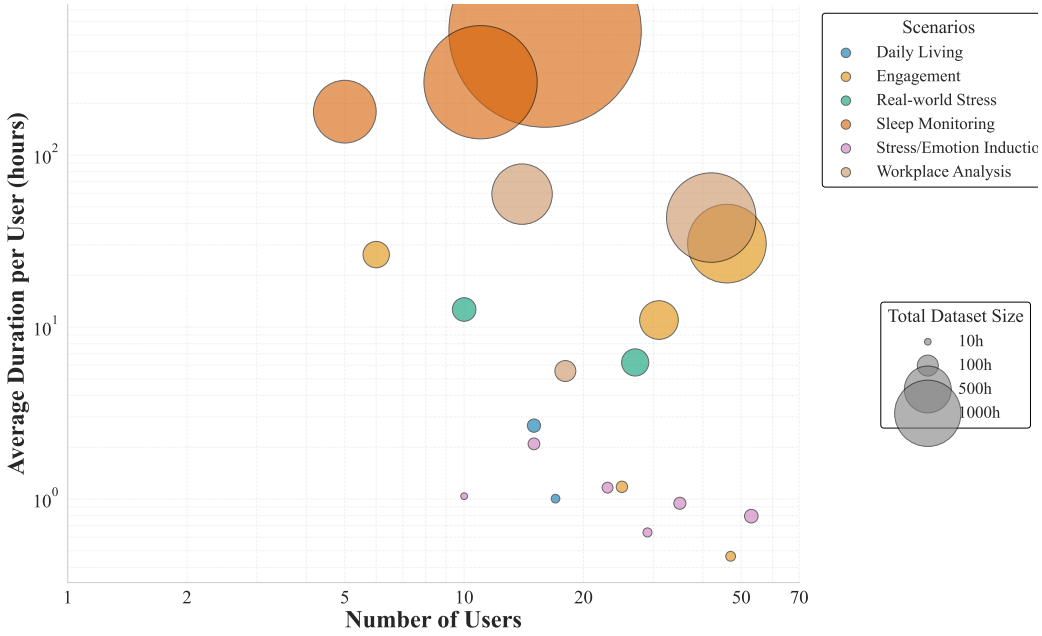


Figure A.3. Number of users and data per user across the EDAMAME collection of datasets. Marker size is proportional to the size of the dataset (in terms of hours of EDA data).

A.3. UME: open source foundation model for EDA data

A.3.1. Data pre-processing and preparation

Data split We divide EDAMAME into two parts: a train and a downstream evaluation parts. For simplicity, we call the first EDAMAME-train and the second EDAMAME-test. We use EDAMAME-train to train our foundation model for wearable EDA data, while EDAMAME-test to perform evaluation on a set of downstream classification tasks. We select 17 datasets for EDAMAME-train and 7 datasets for EDAMAME-test. We perform this split to allow models to be tested on data never seen during training, containing new users, locations, and protocols as well. We select the datasets in the downstream task in order to have a diverse set of labels relevant for EDA data. We report in Table A.3 the binary task from each dataset in the downstream evaluation part.

Pre-processing We pre-process the data from both EDAMAME-train and EDAMAME-test parts following related work on EDA data (Alchieri et al., 2024; Di Lascio et al., 2019; Gashi et al., 2020; Hossain et al., 2022). First, we apply a Butterworth low-pass filter (cutoff 0.4 Hz) to remove high-frequency noise. Then, we decompose the EDA signals into phasic and tonic components, using the cvxEDA method from Greco et al. (2016). This is a common procedure applied when working with EDA data. The phasic component is associated with short-term changes, e.g., momentary stress, while the tonic component with longer term variations (Boucsein, 2012). To train the UME foundation model, we use both the phasic and tonic components, as well as the non-decomposed EDA signal. Using all three signals, i.e., phasic, tonic and original EDA signal, is a common procedure in classification-based tasks using EDA data, since it allows models to learn both short and long-term effects in the data (Alchieri et al., 2024; 2025a).

Data segmentation After pre-processing the data, we segment the EDA signals into fixed-length windows of 60 seconds. We use the same window length as Matton et al. (2023) and Schmidt et al. (2018), who highlight how this length allows to capture both short and long-term changes in the EDA data. To train UME, on EDAMAME-train we select maximum overlapping windows, i.e., with an overlap step of 0.25 s (the sampling rate). We implement this approach as done by Matton et al. (2023), on self-supervised learning for EDA data, and Ansari et al. (2024); Goswami et al. (2024); Feofanov et al. (2025), on generalist time series foundation models for other physiological signals. With this approach, we obtain a train set consisting of approximately *275 million* windows of EDA data.

We also apply the same 60-second segmentation on the data used for evaluation, i.e., EDAMAME-test. For evaluation, we assign to each window a binary label, corresponding to the associated task: we refer to Table A.3 for an overview of the binary downstream tasks used. Whenever a dataset contains EDA collected simultaneously from both sides of the body, for evaluation we use only the data from the body side most associated with the task, following guidelines by Alchieri et al. (2024), e.g., right-side EDA signals for cognitive load classification. On EDAMAME-test, we use non-overlapping windows, since the literature shows how testing on either overlapping or non-overlapping windows leads to similar results (Dehghani et al., 2019; Tello et al., 2024).

Discussion on rescaling the data We do not apply any normalization or standardization, e.g., min-max normalization, on the prepared windows of EDA data. Researchers working with EDA data collected in lab-controlled environments apply per-user min-max normalization to reduce inter-personal variability and improve classification performance significantly, e.g., (Mishra et al., 2020; Schmidt et al., 2018; Almadhor et al., 2023). However, EDA data is affected by both inter-personal variability, which min-max normalization addresses, and intra-personal variability, i.e., a user's data changes over time. A fixed calibration, like the one applied by min-max normalization, is rendered invalid over time if a user's signal morphology changes (Viana-Matesanz & Sánchez-Ávila, 2024; Beten et al., 2025). Moreover, "cold-start" approaches, i.e., a machine learning model is applied directly on a new user, are the preferred approach for wearable physiological data (Yi et al., 2023). Normalization approaches cannot be implemented directly in "cold-start" scenarios.

A.3.2. Model training task & architecture

Training objective We adopt contrastive learning to train UME. Contrastive learning is used by multiple researchers to train foundation models for wearable physiological data, e.g., (Abbaspourazad et al., 2024; Pillai et al., 2025). This approach has also been used for generalist time series foundation models specialized in classification tasks, e.g., (Feofanov et al., 2025). Contrastive learning is a self-supervised learning method that consists in training an encoder-only model to learn a latent embedding space where representations of similar data pairs — typically created via augmentations of the same signal — are attracted to each other, while representations of dissimilar pairs are repelled. The latent embeddings

Table A.3. Summary of Datasets and Associated Binary Tasks

Dataset	Binary Tasks
APSync	Low/High Engagement
HeartS	Sleep/Wake
USILaughs	Cog. Load/Relaxation
WESAD	High/Low Arousal, High/Low Valence (self-report)
Nurses' Stress	Low/High Stress
DREAMT	Sleep/Wake
HHISS	Low/High Stress

often contain information which allows them to be used effectively in downstream tasks (Le-Khac et al., 2020). We also experiment with an additional self-supervised learning method: masked reconstruction. This method consists in training an encoder-decoder to reconstruct the whole signal from a masked version of it, i.e., a signal which is missing some parts. We report in Appendix C details about this additional experiment which, however, failed to learn useful representations of our EDA data.

Model architecture We choose the architecture for our foundation model from similar work in the literature (Abbaspourazad et al., 2024; Narayanswamy et al., 2024; Pillai et al., 2025; Feofanov et al., 2025). Our choice reflects the decision to implement a generalist time series foundation model on EDA data. This choice is in line with research on the first PPG and ECG-specific foundation models (Abbaspourazad et al., 2024; Narayanswamy et al., 2024): our objective is to showcase how features computed from EDA-specific foundation models perform, as well to provide weights and code to the research community. Each physiological signal, from PPG to ECG and EDA, has morphological elements specific to them. It is possible to define architectures for foundation models that adapt to these morphological elements, as recent work from Saha et al. (2025) show. However, this implementation goes beyond the scope of the current work.

We chose an EfficientNet (Tan & Le, 2019) architecture, as Abbaspourazad et al. (2024), since it is computationally less expensive than traditional CNN backbones. We consider this choice also in light of the fact that foundation models for wearable data have the potential to be used on wearable devices themselves (Abbaspourazad et al., 2024). We adapt the EfficientNet for our input data, i.e., 1-dimensional time series data of 240 values (60 s at 4 Hz) with 3 channels (tonic, phasic and original EDA signal). We report in Appendix C ablation studies on the model size and hyperparameters. We implement a version of EfficientNet with approximately 1M parameters and a latent representation of $d = 64$.

A.3.3. Training process

Train loss Using the contrastive learning training objective and EfficientNet architecture defined in **??**, we train UME using the InfoNCE loss (van den Oord et al., 2019). Formally, for a given pair of embeddings (z_i, z_j), the loss is defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, $\tau = 0.1$ is a temperature hyperparameter, N is the batch size, and $\mathbb{1}$ is the indicator function. This objective maximizes the similarity between representations of the same underlying signal while minimizing agreement with unrelated samples.

Pairs of EDA data We generate positive pairs of EDA signals by applying two distinct stochastic augmentations to the same EDA signal segment (an anchor segment). We employ the set of data augmentations optimized for EDA signals proposed by Matton et al. (2023). Negative pairs consist of comparisons between the anchor segment and all other segments in the mini-batch. This set combines both standard data augmentation techniques (e.g., signal warping) and augmentations specific to EDA data (e.g., loose sensor artifact). We report the complete list of augmentations and additional details about the training in Appendix B (Table B.5).

A.4. Evaluation of UME

In this section, we report the results of the evaluation procedure of UME on the selected downstream tasks from EDAMAME-test. We use *linear probing* with frozen weights to evaluate the performance on the selected classification tasks, as frequently done to evaluate performance of foundation models (Abbaspourazad et al., 2024; Saha et al., 2025; Pillai et al., 2025; Zhang et al., 2025). We compare the UME feature set to various baseline features, including generic handcrafted features, a set of EDA-specific features, and features computed from generalist time series foundation models. Finally, we also evaluate the computation complexity of the UME foundation model in extracting features, computed to the other baseline methods.

A.4.1. Experimental evaluation setup

Evaluation through linear probing We evaluate the features computed from the UME foundation model using *linear probing* on the downstream tasks from EDAMAME-test. For each dataset, we freeze the trained weights from the foundation model to compute features on the EDA data. Then, we train a logistic regression classifier on the computed feature set. The usage of frozen weights and linear probing is commonly done in similar work on foundation models for physiological data (Abbaspourazad et al., 2024; Pillai et al., 2025; Saha et al., 2025). We use the same evaluation procedure for UME and the other baseline features sets. We use a linear model for all approaches since we are gauging the representativeness of the features sets on classification tasks, and not that of the downstream model itself. Linear probing also allows to estimate the linear separability of the different feature sets.

Cross-validation protocols for the downstream tasks We evaluate the linear probing using two distinct cross-validation methods. The first method is Leave-One-Participant-Out (LOPO) cross-validation. LOPO cross-validation is used by researchers to evaluate how machine learning models generalize to new users (Rehman et al., 2024). With this method, we test robustness to inter-personal variability. The second validation method is Time-Aware (TA) cross-validation (Alchieri et al., 2025b), which we use to evaluate the model's ability to generalize to data from users already seen in the train set, and this we test robustness to intra-personal variability. We partition users into N folds: the model trains on all external groups plus the first chronological 2/3 of the target fold, reserving the final 1/3 of data strictly for testing. This approach simulates a realistic scenario where a model uses a specific participant's historical data to predict their future states. In our experimental setup, we use $N = 5$.

In both validation methods, we perform hyperparameter tuning for the logistic regression at train set, i.e., at each cross-validation iteration. We perform hyperparameter selection using a 3-fold "inner" cross-validation with grid search. We report in Appendix B information about the hyperparameter grid used.

Baseline feature sets We compare the performance using linear probing on the features computed from UME and using other baseline methods. Specifically, to follow the same experimental setup of similar work on foundation models for physiological data (Abbaspourazad et al., 2024; Pillai et al., 2025; Narayanswamy et al., 2024), we define a set of baseline, *generic*, handcrafted features. These features are: the mean, the standard deviation, the minimum and the maximum of a 60 s EDA signal. We compute these features for all three EDA components, i.e., phasic, tonic and original signal. In total, the dimensionality of this feature set is $d = 12$.

However, the aforementioned generic handcrafted feature set does not represent the state-of-the-art approach for EDA-based classification tasks (Alchieri et al., 2024; Gashi et al., 2020; Lutin et al., 2021). To this end, we also implement a second handcrafted feature set, which we call *EDA-specific* handcrafted features. This feature set includes both statistics, e.g., average of the first derivative, as well as feature specific to EDA data, e.g., number of EDA peaks and their average amplitude. As with the generic handcrafted feature set, we compute these features on 60 s windows and for all three EDA components. In total, the EDA-specific handcrafted features are $d = 45$.

We also compare the performance using linear proving of the feature set from UME with other generalist time series foundation models. We compare with the following: Chronos (Ansari et al., 2024), MOMENT (Goswami et al., 2024) and Mantis (Feofanov et al., 2025). We select these three foundation models since Alchieri et al. (2025a) show how they achieve performance similar to the EDA-specific handcrafted features when using EDA data. Saha et al. (2025) also show that Chronos (Ansari et al., 2024) and MOMENT (Goswami et al., 2024) achieve, on average, performance similar to their PPG-specific foundation model on the downstream tasks selected by the authors. We also select the recent foundation model Mantis (Feofanov et al., 2025) since it is trained specifically for classification tasks on time series data.

Mantis takes time series of length 512 as input: we oversample our time series, which have length of 240, to match this

desired length. From a single time series, Mantis computes a set of embeddings, similarly to our foundation model. Mantis has a feature set size of $d = 768$ (embedding size of 256 across three channels). The embedding size of MOMENT is $d = 1024$. However, Chronos computes embeddings ($d = 512$) for each timepoint in the series. This leads, with our input, to high dimensionality of the data. To address the high dimensionality problem, we average across the time axis the embeddings extracted from each timepoint. Overall, the feature set from Chronos is of size $d = 1536$ (embedding size of 512 across three channels).

Reporting and task selection information We report all results in terms of *balanced accuracy*. We use this metric since the downstream evaluation datasets we use contain binary labels which, in a subset of cases, are imbalanced. Balanced accuracy accounts for class imbalance, reporting results which are more representative of the real performance on a specific task (Brodersen et al., 2010; Owusu-Adjei et al., 2023). At the same time, compared to other metrics for imbalanced data, e.g., Matthew's correlation coefficient (MCC), balanced accuracy is more interpretable (Grandini et al., 2020).

In addition to the results from the identified feature sets, we also report results from a dummy classifier. The dummy classifier reported is the one achieving the highest balanced accuracy in each given task, among the following: *most frequent*, which predicts always the most frequent class set; *uniform*, which predicts the binary labels by drawing randomly from a uniform distribution; and *prior*, which predicts the binary labels by drawing randomly from the distribution of labels in the train set.

We report in this section results from tasks that are *solvable*, i.e., at least one feature set achieves balanced accuracy higher than the dummy classifier. If no model achieves balanced accuracy higher than, for example, random chance, we conclude that it is not due to issues with the feature sets, but with the task itself, e.g., the task is not solvable with the given constraints.

A.4.2. Comparison with generic handcrafted feature set

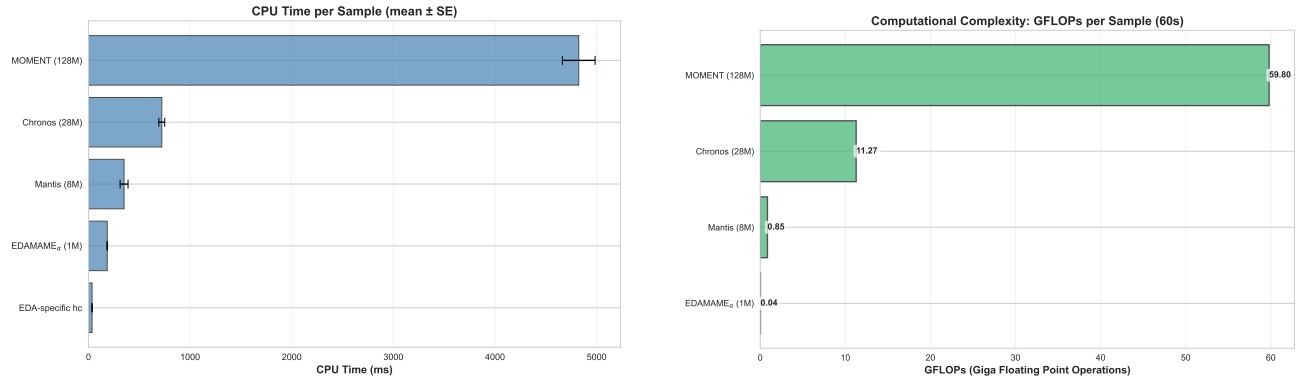
We present in Figure 2 the comparison when performing linear probing on the selected downstream tasks, between the feature set from our UME foundation model and the generic handcrafted features. We show results for both validation methods, i.e., LOPO and TA. The results show that the features from our foundation model outperform the generic handcrafted features in 9 out of 10 tasks. The improvement holds true regardless of the scenario and the task selected. We conclude from these findings that the UME foundation model learns features useful for EDA data in performing downstream predictions. Our findings are in line with works on foundation models for PPG, when comparing to generic handcrafted feature sets (Abbaspourazad et al., 2024; Pillai et al., 2025; Narayanswamy et al., 2024).

A.4.3. Comparison with other baseline feature sets

We report in Table A.4 the results for the evaluation of our UME foundation model, in terms of balanced accuracy and for two distinct validation methods, i.e., Time-Aware (TA) cross-validation and Leave-One-Participant-Out (LOPO) cross-validation. We compare the feature set computed using UME to the following feature sets: a set of generic handcrafted features; a set of EDA-specific handcrafted features; and features computed using generalist time series foundation models, specifically Mantis (Feofanov et al., 2025), Chronos (Ansari et al., 2024) and MOMENT (Goswami et al., 2024).

Results using TA The TA method tests the intra-personal generalizability of models trained using the different feature sets, i.e., the ability to generalize to new data from users already seen in the train data. Our first finding is that, using linear probing, models trained using features from UME always outperform non-foundation model methods, i.e., both generic and EDA-specific handcrafted features. We conclude that the embeddings from UME capture user-specific dynamics, which allow to obtain a balanced accuracy on par, or higher, than handcrafted-based methods. We also find that models trained using the feature set from Mantis (Feofanov et al., 2025), which is trained to solve generic classification tasks, perform similarly to those trained using the feature set from UME. We notice how the other two foundation models, MOMENT (Goswami et al., 2024) and Chronos (Ansari et al., 2024), have lower performance, on a majority of tasks, than both the two handcrafted feature sets and the foundation models (UME and Mantis (Feofanov et al., 2025)).

Results using LOPO The LOPO cross-validation method tests the inter-personal generalizability of models trained using the different feature sets, i.e., the ability to generalize to data from new users not seen in the train set. We find that models trained from the UME features have similar balanced accuracy to models trained from either the EDA-specific handcrafted features and the embeddings computed using Mantis (Feofanov et al., 2025). Features from the other foundation models have similar or lower performance as well. Generalization to new users is a known issue when working with EDA data (Gashi



(a) CPU Time *(b)* Model GFLOPs

Figure A.4. Computational complexity analysis. **(a)** Average (across 20 samples) per-sample computation time for a 60-second window of EDA data for all three components (phasic, tonic, and original signal). Note: *hc* stands for *handcrafted*; generic handcrafted features are not shown as they require negligible computation time. **(b)** GFLOPs for the different foundation models.

et al., 2020). While features from foundation models, both our model and others, achieve performance similar to that of EDA-specific handcrafted features, they do not solve the "cold-start" problem.

Additional remarks on standard errors We highlight how, regardless of the feature set used, the standard error associated with all results leads to overlapping confidence intervals. In other words, while there are trends, e.g., models trained using features from UME outperforming models trained using EDA-specific handcrafted features in a majority of tasks, there is no statistical difference between results obtained using the different feature sets. However, all results are statistically higher (t-test corrected with Bonferroni) compared to the results obtained using the dummy classifier.

Result analysis using Friedman-Nemenyi test We perform the Friedman test, followed by the post-hoc Nemenyi test, to compare model performance across all experiments (Demšar, 2006). We consider each combination of dataset-validation method as a single sample for the statistical analysis. The Friedman test is used to determine whether *any* statistical difference is present across all models and experiments. For the Friedman test, we find a p-value of approximately 0.0001, which is below the reference threshold of $\alpha = 0.05$. We conclude from this result that, across all experiments, there are statistically significant differences. In particular, we attribute this finding to the difference between all methods, i.e., both handcrafted- and foundation model-based, and the dummy classifier baseline. We use the post-hoc Nemenyi test to perform pair-wise statistical comparisons, from the model rankings provided by the Nemenyi test. We report the findings with respect to our UME foundation model only. We refer to the Appendix C to additional results. First, we find that our method achieves performance statistically higher than the dummy classifier baseline ($p = 0.02 < \alpha = 0.05$). Secondly, we find that the performance difference in linear probing between feature sets computed using UME and Mantis (Feofanov et al., 2025) is not statistically different ($p \simeq 0.9 > \alpha > 0.05$). Finally, we also find no statistical difference between using our UME and the EDA-specific handcrafted features ($p \simeq 0.9 > \alpha = 0.05$).

From the Friedman-Nemenyi test findings, we conclude that both our UME and Mantis (Feofanov et al., 2025) capture time series dynamics which allow to achieve, on average, performance similar to the EDA-specific handcrafted feature sets. Our EDA-trained foundation model performs on-par with a large-scale generalist time series foundation model, Mantis (Feofanov et al., 2025), even if trained on a relatively smaller dataset and with fewer parameters (1M ours vs 8M Mantis).

A.4.4. Computational complexity analysis

We compare the computational complexity of the different feature extraction methods, i.e., the handcrafted-based approaches, our UME foundation model, and the other baseline foundation models. We compare the methods using both *CPU execution time* and FLOPs (Chen et al., 2023). We choose this dual approach since handcrafted features cannot be compared using FLOPs alone. We perform all calculations using an Apple M1 Max CPU.

To get *CPU execution time*, we use a random subset of twenty 60 s windows from the downstream part of the UME collection. We also use additionally three samples per experiment as warm-up runs. We perform FLOPs computation using

Table A.4. Results of binary classification experiments across datasets and tasks, in terms of *balanced accuracy*$_{standard\ error}$. Reported are results for both TA and LOPO cross-validation methods. Acronyms: *Gen. HC* stands for *generic handcrafted features*; *EDA HC* stands for *EDA-specific handcrafted features*.

Dataset	Binary Task	Dummy	Handcrafted (HC)		Generalist Foundation			Ours
			Gen.	EDA-spec.	Mantis	MOMENT	Chronos	UME
		Balanced accuracy$_{standard\ error}$						
TIME-AWARE CROSS-VALIDATION (TA)								
APSYNC	Low/High engagement	$.45_{.20}$	$.80_{.10}$	$.80_{.10}$	$.86_{.07}$	$.47_{16}$	$.61_{.20}$	$.89_{.11}$
Dreamt	Deep Sleep/REM	$.48_{.02}$	$.55_{.05}$	$.59_{.04}$	$.64_{.03}$	$.65_{01}$	$.63_{02}$	$.63_{.05}$
Dreamt	Sleep/Wake	$.50_{.01}$	$.65_{.02}$	$.69_{.01}$	$.73_{.01}$	$.69_{01}$	$.73_{01}$	$.70_{.01}$
HeartS	Sleep/Wake	$.49_{.00}$	$.66_{.04}$	$.72_{.07}$	$.75_{.09}$	$.69_{06}$	$.74_{06}$	$.73_{.09}$
WESAD	Low/High Valence	$.51_{.03}$	$.55_{.10}$	$.54_{.10}$	$.61_{.04}$	$.64_{05}$	$.45_{.06}$	$.63_{.06}$
LEAVE-ONE-PARTICIPANT-OUT (LOPO) CROSS-VALIDATION								
Dreamt	Sleep/Wake	$.48_{.00}$	$.70_{.01}$	$.74_{.01}$	$.76_{01}$	$.73_{01}$	$.78_{01}$	$.75_{.01}$
USILaughs	Cog. load/relax	$.50_{.00}$	$.66_{.03}$	$.70_{.04}$	$.72_{.05}$	$.60_{.05}$	$.67_{.03}$	$.71_{.05}$
HeartS	Sleep/Wake	$.50_{.00}$	$.66_{.04}$	$.70_{.03}$	$.74_{.03}$	$.70_{.02}$	$.73_{.03}$	$.72_{.03}$
WESAD	Low/High Arousal	$.58_{.05}$	$.61_{.04}$	$.63_{.04}$	$.56_{.04}$	$.66_{.04}$	$.66_{.05}$	$.61_{.05}$
HHISS	Stress/calm	$.50_{.00}$	$.56_{.02}$	$.64_{.01}$	$.63_{.02}$	$.55_{.01}$	$.59_{.01}$	$.60_{.02}$

the `fvcore` Python library. We report in Figure A.4 the results. We find that the handcrafted feature extraction methods have the lowest CPU execution time, as expected. We also find that our UME foundation model is significantly faster than all other foundation models, both with respect to *CPU execution time* and FLOPs. We conclude that our model, which is trained specifically only on EDA data, achieves performance on-par with larger, general purpose, foundation models on a set of downstream tasks using EDA data, as detailed in Appendix A.4.3. However, it does so with a fraction of the computation resources needed, while also eliminating the need for expert-designed or handcrafted features, unlike traditional feature-based pipelines

A.5. Dataset and code availability

We share the code to prepare and pre-process the data as we did at the following anonymized link: https://anonymous.4open.science/r/datasets_cleaning-E7F9.

We also make the code to train and evaluate the UME foundation model available as open source to other researchers, as well as the model weights. For the review process, we share an anonymized version of the code to train and the weights at the following link: https://anonymous.4open.science/r/eda-foundation-models-5BBB.

Moreover, we also share the code used to validate UME and the other feature extraction methods on the downstream tasks at the following anonymized link: https://anonymous.4open.science/r/pretrained_foundation_models_physiological_data-5CB3.

Table B.1. Summary of datasets with their corresponding data-sharing agreements. This table includes all datasets, including citations and the "Reshareable" column, for inclusion in the appendix.

Dataset Name	License
BiHeartS (Abdalazim et al., 2025)	✓- data share agreement
BIG IDEAS (Bent et al., 2021)	✓- with original license (ODC-BY 1)
Dynamics in the workplace (Lukan et al., 2018)	✓- data share agreement
EmpaticaE4Stress (Campanella et al., 2023)	✓- (CCBY 4) with original license
EPM-E4 (Garcia-Moreno & Badenes-Sastre, 2020)	✓- (CCBY 4) with original license
LAUREATE (Laporte et al., 2023)	✓- data share agreement
MEFAR (Derdiyok et al., 2024)	✓- (CCBY 4) with original license
M2Sleep (Gashi et al., 2022a)	✓- data share agreement
PPG-Dalia (Reiss et al., 2019)	✓- (CCBY 4) with original license
SEED (Di Lascio et al., 2018)	✓- data share agreement
SEED-II-Lab (Di Lascio et al., 2018)	✓- data share agreement
SEED-II-Wild (Di Lascio et al., 2018)	✓- data share agreement
Stress Predict (Iqbal et al., 2022)	✓- (MIT) with original license
ToadStool (Svoren et al., 2020)	✓- (CCBY 4) with original license
WEEE (Gashi et al., 2022b)	✓- (CCBY 4) with original license
WESD (Rafiul Amin et al., 2022)	✓- with original license (ODC-BY 1)
Workplace (Di Lascio et al., 2021)	✓- data share agreement
APSync (Gashi et al., 2019)	✓- data share agreement
HeartS (Abdalazim et al., 2023)	✓- data share agreement
USILaughs (Di Lascio et al., 2019)	✓- data share agreement
WESAD (Schmidt et al., 2018)	✓- (CCBY 4) with original license
Nurses' Stress (Hosseini et al., 2022)	✓- (ODbL 1) with original license
DREAMT (Wang et al., 2024)	✗- PhysioNet Restricted Health Data Use Agreement 1.5.0
HHISS (Gjoreski et al., 2020)	✗- no license specified

B. Additional information for reproducibility

B.1. Additional dataset information

We report in Table B.1 information about the original license associated with the datasets making up the EDAMAME collection. The datasets that we re-share are all under their original license, if so required. For the datasets that require a data sharing agreement to be signed, we contacted the original authors, which all agreed to for re-sharing under the conditions in the original data sharing agreement.

B.2. Baseline feature sets

In Appendix A.4.1 we explain how we use two handcrafted feature sets as baseline to evaluate the embeddings from our UME foundation model. The first feature set, which we call *generic handcrafted feature set*, consists of four statistics, i.e., mean, minimum, maximum and standard deviation, computed for both the phasic and tonic EDA components, as well as the original non-decomposed EDA signal. This feature set emulates baseline used by related work to evaluate foundation models for physiological data (Abbaspourazad et al., 2024; Pillai et al., 2025). The second set of features, which we call *EDA-specific handcrafted feature set*, is a broader set of features commonly used in EDA classification pipelines (Alchieri et al., 2024; Gashi et al., 2020). In addition to the generic statistics, this set incorporates signal dynamics (e.g., slopes and derivatives), morphological peak characteristics, and frequency-domain components. This set consists of 15 base features, resulting in a total of 45 features per window (15 features $\times$ 3 EDA components).

Table B.3 details the exact mathematical formulations for all 15 base features and specifies their inclusion in the respective baseline sets.

We report in Table B.2 the feature size of the feature extraction methods used in this work.

B.3. Hyperparameter grid for the logistic regression

In Table B.4 we report the hyperparameter grid we used with the logistic regression. We used a 3-fold inner cross-validation to search for the best hyperparameter configuration.

Table B.2. Overview of the models and feature sets used in the experimental evaluation. We report the model size (number of trainable parameters) and the dimensionality of the feature space (d) used for linear probing.

Model / Feature Set	Model Size	Feature Dim. (d)
HANDCRAFTED METHODS		
Generic HC	N/A	12
EDA-specific HC	N/A	45
GENERALIST FOUNDATION MODELS		
Mantis (Feofanov et al., 2025)	$\sim$ 8 M	768
Chronos (Ansari et al., 2024)	$\sim$ 200 M	1536
MOMENT (Goswami et al., 2024)	$\sim$ 385 M	1024
OURS		
UME	$\sim$ **1 M**	**64**

B.4. Training details

We train the UME foundation model using the Adam optimizer (Kingma & Ba, 2017) (learning rate 0.001 and weight decay 0.01). We also use a learning rate scheduler, Reduce On Plateau (factor 0.5), to decrease the learning rate during training. We train the model for a maximum of 400 epochs, with early stopping, with a batch size of 512. In total, we train our model for approximately 5 days, using an Nvidia A6000 GPU[13]. We implement the foundation model in Python, using the Pytorch (Paszke et al., 2019) and Pytorch-Lightning libraries.

B.5. Data augmentations for EDA data

We report in Table B.5 the list of data augmentations used to train our UME foundation model with contrastive learning. We use the same set proposed by Matton et al. (2023), which contains EDA-specific augmentations.

[13] https://www.nvidia.com/en-us/products/workstations/rtx-a6000/

Table B.3. Mathematical formulations for the 15 base hand-crafted features extracted from EDA time series signals (x) of length N. All features are computed independently for the tonic, phasic, and original EDA components. The "Feature Set(s)" column indicates whether the feature was used in the Generic baseline (12 total features) or the EDA-specific baseline (45 total features).

Feature	Mathematical Notation or Formula	Feature Set(s)
TIME-DOMAIN		
Mean	$\frac{1}{N}\sum_{i=1}^{N} x_i$	Generic & EDA-specific
Minimum	$\min(x_1, x_2, \ldots, x_N)$	Generic & EDA-specific
Maximum	$\max(x_1, x_2, \ldots, x_N)$	Generic & EDA-specific
Standard Deviation	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$	Generic & EDA-specific
Dynamic Range	$\max(x_1, x_2, \ldots, x_N) - \min(x_1, x_2, \ldots, x_N)$	EDA-specific only
Slope	$\frac{x_N - x_1}{N-1}$	EDA-specific only
Absolute Value of Slope	$\left\lvert\frac{x_N - x_1}{N-1}\right\rvert$	EDA-specific only
Mean of the First Derivative	$\frac{1}{N-1}\sum_{i=1}^{N-1}(x_{i+1} - x_i)$	EDA-specific only
Std. Dev. of the First Derivative	$\sqrt{\frac{1}{N-2}\sum_{i=1}^{N-1}((x_{i+1} - x_i) - \bar{x'})^2}$	EDA-specific only
Number of EDA Peaks	Count of local maxima in the window	EDA-specific only
Amplitude of EDA Peaks	Mean amplitude of local maxima in the window	EDA-specific only
FREQUENCY DOMAIN (FAST FOURIER TRANSFORM)		
Direct Current (DC)	X_0	EDA-specific only
Sum of Frequency Coefficients	$\sum_{k=1}^{N} \lvert X_k \rvert$	EDA-specific only
Information Entropy	$-\sum_{k=1}^{N} P(X_k)\log_2(P(X_k))$	EDA-specific only
Spectral Energy	$\sum_{k=1}^{N} \lvert X_k \rvert^2$	EDA-specific only

Table B.4. Hyperparameter Grid for Logistic Regression

Hyperparameter	Values
Regularization strength (C)	0.01, 0.1, 1, 10
Solver	lbfgs, liblinear
Penalty	L2
Max iterations	10000
Class weight	balanced

Table B.5. Summary of Electrodermal Activity (EDA) Data Augmentations used to train the UME foundation model. The table details both EDA-specific transforms designed to isolate physiological components or simulate artifacts, and generic time series transforms. The set is the same proposed by Matton et al. (2023).

Augmentation	Type	Description	Parameter Range
FREQUENCY DOMAIN & COMPONENT ISOLATION			
Low-Pass Filter	EDA-Specific	Extracts tonic component; removes high-freq noise.	Cutoff $f \in [0.25, 1.0]$ Hz
High-Pass Filter	EDA-Specific	Extracts phasic component; removes slow drifts.	Cutoff $f \in [0.05, 0.25]$ Hz
Band-Pass Filter	EDA-Specific	Isolates information-rich EDA frequency bands.	Pass $f \in [0.05, 0.25]$ Hz
Band-Stop Filter	EDA-Specific	Rejects specific frequency bands.	Reject $f \in [0.75, 1.0]$ Hz
High Freq. Noise	EDA-Specific	Adds Gaussian noise only to frequencies $>$ 1Hz.	Noise $\sigma \in [0, 0.5]$
ARTIFACT SIMULATION			
Jump Artifact	EDA-Specific	Simulates abrupt sensor movement/displacement.	Jump $\in [0.01, 0.2]\mu S$
Loose Sensor	EDA-Specific	Simulates electrode contact loss (signal drop).	Duration $t \in [5, 20]$ s
THERMOREGULATION SIMULATION			
Tonic Const. Scale	EDA-Specific	Scales tonic component (simulates constant temp).	Factor $s \in [0.25, 2]$
Tonic Amp. Warp	EDA-Specific	Time-varying scale of tonic component (changing conditions).	Spline $\sigma \in [0.01, 0.05]$
GENERIC TIME SERIES			
Amp. Const. Scale	Generic	Applies constant scaling factor to the raw signal.	Factor $s \in [0.25, 2]$
Amplitude Warp	Generic	Applies smooth, time-varying scaling to raw signal.	Spline $\sigma \in [0.01, 0.05]$
Gaussian Noise	Generic	Adds random Gaussian noise to the raw signal.	$\sigma \in [0, 0.5]$
Time Shift	Generic	Shifts the signal window forward or backward.	Shift $t \in [5, 45]$ s
Temporal Cutout	Generic	Masks/zeroes out a random sub-window of the signal.	Cutout $t \in [5, 15]$ s
Time Warp	Generic	Perturbs temporal dimension (local stretch/compress).	Spline $\sigma \in [0.01, 0.1]$
Permutation	Generic	Slices signal and randomly reorders sub-windows.	Segments $n \in [2, 6]$
Signal Flip	Generic	Inverts the signal over its amplitude dimension.	N/A

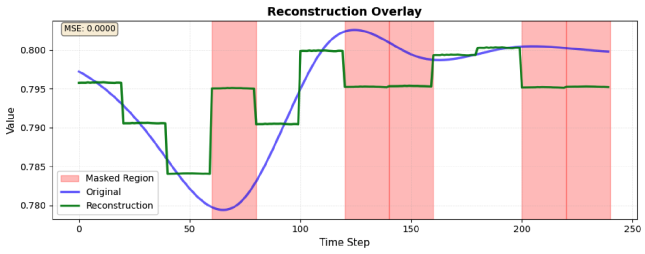


Figure C.1. Example of EDA signal reconstruction using MAE.

C. Sensitivity & additional model studies

We perform sensitivity and additional studies to find the configuration of our UME foundation model. In this section, we report details for: implementation of an additional NN architecture, i.e., a Masked Autoencoder (MAE); studies on model size and hyperparameters for the EfficientNet architecture chosen for the UME foundation model, trained using contrastive learning.

C.1. Experiments with Masked Autoencoders (MAEs)

We train a reconstructive-based masked autoencoder (MAE), based on vision transformer (ViT) (Dosovitskiy et al., 2020). Narayanswamy et al. (2024) also used a masked-autoencoder to train a foundation model for physiological data. We train the model using the following loss: let $\mathbf{x} \in \mathbb{R}^{T \times C}$ denote the input EDA signal, partitioned into N non-overlapping patches of size P, and let $\hat{\mathbf{x}}_i$ and $\mathbf{x}_i$ denote the reconstructed and original i-th patch, respectively. Given a binary mask $\mathbf{m} \in \{0,1\}^N$, where $m_i = 1$ indicates a masked patch and $m_i = 0$ a visible one, the reconstruction loss is defined as

$$\mathcal{L} = \alpha\, \mathcal{L}_{\text{masked}} + (1-\alpha)\, \mathcal{L}_{\text{visible}}, \tag{2}$$

with

$$\mathcal{L}_{\text{masked}} = \frac{1}{\sum_i m_i} \sum_{i=1}^{N} m_i\, \ell(\hat{\mathbf{x}}_i, \mathbf{x}_i), \tag{3}$$

$$\mathcal{L}_{\text{visible}} = \frac{1}{\sum_i (1-m_i)} \sum_{i=1}^{N} (1-m_i)\, \ell(\hat{\mathbf{x}}_i, \mathbf{x}_i), \tag{4}$$

where $\ell(\cdot,\cdot)$ denotes either the mean absolute error (MAE) computed within each patch, and $\alpha \in [0,1]$ controls the relative importance of masked versus visible patch reconstruction.

We experiment with the following configurations: $m \in \{0, 0.1, 0.4\}$ and $\alpha \in \{0.1, 0.5\}$. In Figure C.2 we report the validation loss during training curve for the configurations tested, and in Figure C.1 we show an example of a signal reconstruction at train end, over a validation sample. We also perform downstream evaluation on the BiHeartS dataset: however, no configuration leads to performance, in terms of balanced accuracy, above that of the Dummy classifier.

Given these findings, we believe that masked autoencoders are not optimal for the specific EDA data present in the EDAMAME collection of datasets.

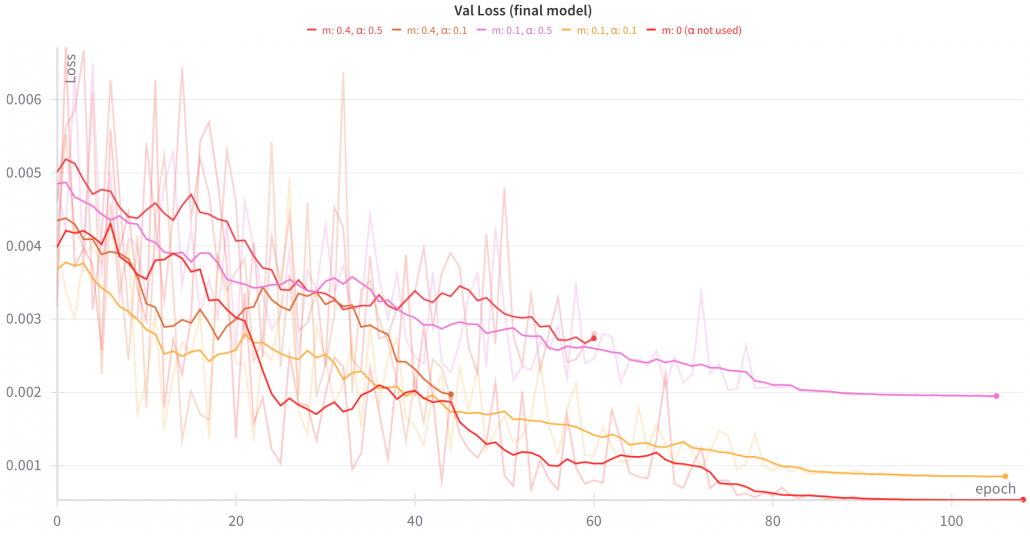


Figure C.2. Validation loss during training for different configurations of masking and masking weight.

Table C.1. Hyperparameter configurations for the four EfficientNet model variants. Across all versions, several hyperparameters remain fixed, including the number of input channels ($C_{in} = 3$), the input sequence/segment length ($L = 240$), the base convolution stride ($S = 2$), and the use of the Swish activation function.

Hyperparameter	Tiny	Small	Base	Large
Stem Channels	32	48	64	96
MBConv Channels	32	48	64	96
MBConv Blocks	4	8	16	24
Head Channels	96	160	248	384
Embedding Dim	32	48	64	96
Kernel Size	5	7	9	11
Dropout	0.3	0.4	0.5	0.5

Table C.2. Sensitivity analysis of UME model size under leave-one-participant-out (LOPO) cross-validation.

Dataset	Task	Random	Gen.	EDA	Tiny	Small	Base	Large
USILaughs	Cog. load / relax	50_{00}	66_{03}	70_{04}	61_{03}	64_{03}	71_{05}	71_{05}

C.2. Sensitivity study on model size

We adopt an EfficientNet architecture for our UME foundation model, which we train using contrastive learning on EDA data. We perform a sensitivity study on the model size. To this end, we train 4 versions of the UME foundation model: tiny ($\sim 50k$ parameters), small ($\sim 200k$ parameters), base ($\sim 750k$ parameters) and large ($\sim 2M$ parameters). We report the details of their individual hyperparameters in Table C.1. We pretrain all of these models using the complete pretrain part of the EDAMAME collection, as in the final model described in **??**; we also use the same pretrain configurations, e.g., number of epochs, as the final model, as reported in Appendix A.3.3.

In order to avoid bias in the result presentation, we compare the performance of these different configurations using only two downstream tasks: sleep/wake from the BiHeartS, which is an in-distribution dataset; and cognitive load/relaxation from the USILaughs, which is out-of-distribution from the pretrain data and a task relevant for EDA.

We report the results in Table C.2. We notice that, across both datasets, the larger the model the higher the balanced accuracy in both tasks. However, we also notice that in the in-distribution dataset this difference is relatively small compared to the other task. We can attribute this to the fact that the models had already seen the data during pretraining. We also notice that the difference between the *base* and *large* models are none, suggesting that increasing beyond the base model size, in our pretrain data, does not lead to an increase in results.

C.3. Sensitivity study on dataset size

Referred to as *foundation model scaling laws*, researchers have shown how the size of the pretrain dataset plays a significant role in the downstream performance of foundation model (Narayanswamy et al., 2024; Shuai et al., 2026). To this end, we perform a sensitity study. In particular, we consider the base UME model and pretrain in on a different subsets of the EDAMAME-train subcollection: 5%, 20%, 50% and 75%.

Similarly to the study in Appendix C.2, we compare the performance of the different configurations using a single downstream task, cognitive load/relaxation on the USILaughs dataset. We use the same pretrain setup.

We report the results in Table C.3, in terms of balanced accuracy. We notice that our foundation model follows the foundation model scaling laws as similar work from the literature, e.g., (Narayanswamy et al., 2024; Shuai et al., 2026).

From these results we also conclude that the size of the pretrain dataset, EDAMAME, is indeed necessary to achieve results higher than existing approaches.

Table C.3. Results for *sensitivity analysis of dataset size on UME training*. Experiments were run using the base UME model. Random undersampling of the complete pretrain EDAMAME dataset was used. Reported is the *balanced accuracy*.

Dataset	**Binary Task**	**Random**	**Gen.**	**EDA-spec.**	**5%**	**20%**	**50%**	**75%**	**UME (100%)**
USILaughs	Cog. load/relax	50_{00}	66_{03}	70_{04}	61_{03}	63_{05}	64_{03}	68_{04}	71_{05}

Table D.1. CPU Performance Comparison of Extractors

Extractor	CPU Time (mean) [ms]	CPU Time (SE) [ms]
MOMENT (128M)	4823.53	161.11
Chronos (28M)	721.34	28.91
Mantis (8M)	350.23	39.06
UME (1M)	184.56	3.12
EDA-specific hc	35.39	1.40
generic hc	0.09	0.01

D. Additional results

Computational analysis complexity We report in Table D.1 the detailed performance metrics for the computation complexity analysis. This table mirrors the results in Figure A.4.

Complete results for the downstream evaluation We report in Table D.2 and Table D.3 the complete results for our experimental evaluation. Specifically, we report the results in terms of balanced accuracy, Matthew's Correlation Coefficient (MCC) (Chicco & Jurman, 2020) and F1-score. We report the results for both validation methods used, i.e., LOPO and TA cross-validations. Results are reported for all tasks in both validation methods. For the TA cross-validation, we do not report the results for *cognitive-load/relaxation* classification on the USILaughs dataset (Di Lascio et al., 2019), since there was not enough data for each user to perform the temporal split.

Table D.2. Results of binary classification experiments, using the Leave One Participant Out (LOPO) cross-validation. Results shown as `metric ± standard error`.

Dataset	Binary Task	Model	Metrics		
			Balanced Acc.	MCC	F1
Dreamt	Deep Sleep/REM	Dummy	0.73 ± 0.05	-0.01 ± 0.01	0.16 ± 0.06
		Gen. HC	0.64 ± 0.07	0.00 ± 0.03	0.10 ± 0.06
		EDA-spec. HC	0.54 ± 0.05	0.02 ± 0.03	0.19 ± 0.07
		Mantis	0.59 ± 0.05	0.02 ± 0.04	0.18 ± 0.07
		MOMENT	0.64 ± 0.03	0.06 ± 0.03	0.21 ± 0.07
		Chronos	0.64 ± 0.04	0.03 ± 0.03	0.19 ± 0.07
		UME	0.56 ± 0.05	0.04 ± 0.04	0.17 ± 0.06
Dreamt	Sleep/Wake	Dummy	0.50 ± 0.00	0.00 ± 0.00	0.49 ± 0.01
		Gen. HC	0.70 ± 0.03	0.40 ± 0.05	0.68 ± 0.04
		EDA-spec. HC	0.74 ± 0.02	0.48 ± 0.04	0.72 ± 0.03
		Mantis	0.77 ± 0.02	0.52 ± 0.04	0.74 ± 0.03
		MOMENT	0.72 ± 0.02	0.44 ± 0.03	0.71 ± 0.03
		Chronos	0.78 ± 0.02	0.55 ± 0.03	0.75 ± 0.03
		UME	0.75 ± 0.02	0.50 ± 0.04	0.73 ± 0.03
HHISS	Stress/calm	Dummy	0.50 ± 0.00	-0.00 ± 0.01	0.24 ± 0.03
		Gen. HC	0.56 ± 0.03	0.12 ± 0.07	0.44 ± 0.06
		EDA-spec. HC	0.64 ± 0.03	0.28 ± 0.07	0.55 ± 0.04
		Mantis	0.63 ± 0.03	0.26 ± 0.06	0.53 ± 0.04
		MOMENT	0.55 ± 0.02	0.09 ± 0.05	0.44 ± 0.03
		Chronos	0.59 ± 0.03	0.17 ± 0.06	0.48 ± 0.04
		UME	0.59 ± 0.03	0.19 ± 0.06	0.49 ± 0.05
HeartS	Sleep/Wake	Dummy	0.50 ± 0.00	0.00 ± 0.00	0.08 ± 0.02
		Gen. HC	0.66 ± 0.09	0.28 ± 0.16	0.32 ± 0.11
		EDA-spec. HC	0.70 ± 0.05	0.32 ± 0.13	0.36 ± 0.11
		Mantis	0.74 ± 0.06	0.37 ± 0.14	0.41 ± 0.13
		MOMENT	0.70 ± 0.04	0.27 ± 0.06	0.33 ± 0.09
		Chronos	0.73 ± 0.06	0.34 ± 0.11	0.39 ± 0.12
		UME	0.71 ± 0.06	0.30 ± 0.10	0.35 ± 0.10
WESAD	Low/High Arousal	Dummy	0.58 ± 0.05	-0.02 ± 0.03	0.05 ± 0.03
		Gen. HC	0.61 ± 0.09	0.10 ± 0.14	0.26 ± 0.13
		EDA-spec. HC	0.63 ± 0.08	0.13 ± 0.11	0.33 ± 0.13
		Mantis	0.56 ± 0.08	0.06 ± 0.13	0.27 ± 0.14
		MOMENT	0.51 ± 0.06	0.03 ± 0.10	0.22 ± 0.12
		Chronos	0.56 ± 0.09	0.07 ± 0.14	0.28 ± 0.10
		UME	0.58 ± 0.09	0.03 ± 0.13	0.26 ± 0.12
WESAD	Low/High Valence	Dummy	0.51 ± 0.06	0.07 ± 0.07	0.80 ± 0.10
		Gen. HC	0.55 ± 0.15	0.01 ± 0.17	0.66 ± 0.18
		EDA-spec. HC	0.58 ± 0.13	0.10 ± 0.13	0.72 ± 0.17
		Mantis	0.55 ± 0.13	0.07 ± 0.16	0.76 ± 0.16
		MOMENT	0.63 ± 0.07	0.03 ± 0.07	0.76 ± 0.14
		Chronos	0.56 ± 0.10	0.05 ± 0.11	0.77 ± 0.15
		UME	0.59 ± 0.13	0.08 ± 0.12	0.77 ± 0.13
APSYNC	Low/High engagement	Dummy	0.46 ± 0.06	-0.09 ± 0.11	0.29 ± 0.16
		Gen. HC	0.61 ± 0.13	0.14 ± 0.25	0.53 ± 0.26
		EDA-spec. HC	0.63 ± 0.16	0.18 ± 0.31	0.52 ± 0.27
		Mantis	0.62 ± 0.11	0.30 ± 0.19	0.48 ± 0.29
		MOMENT	0.60 ± 0.20	0.24 ± 0.39	0.61 ± 0.21
		Chronos	0.48 ± 0.16	-0.02 ± 0.31	0.43 ± 0.25
		UME	0.54 ± 0.13	0.10 ± 0.29	0.49 ± 0.30
USILaughs	Cog. load/relax	Dummy	0.50 ± 0.00	0.00 ± 0.00	0.80 ± 0.00
		Gen. HC	0.66 ± 0.06	0.32 ± 0.13	0.72 ± 0.04
		EDA-spec. HC	0.70 ± 0.09	0.40 ± 0.19	0.70 ± 0.11
		Mantis	0.72 ± 0.09	0.43 ± 0.19	0.76 ± 0.10
		MOMENT	0.60 ± 0.09	0.20 ± 0.18	0.61 ± 0.11
		Chronos	0.68 ± 0.07	0.35 ± 0.14	0.57 ± 0.12
		UME	0.71 ± 0.11	0.42 ± 0.22	0.76 ± 0.10

Table D.3. Results of binary classification experiments for Time-Aware (TA) cross-validation. Results shown as `metric ± standard error`. Experiments whose result is reported as OoM (out of memory) means that the required memory size for the embeddings computation exceeded the computational resources available.

Dataset	Binary Task	Model	Metrics		
			Balanced Acc.	MCC	F1
Dreamt	Deep Sleep/REM	Dummy	0.51 ± 0.01	0.01 ± 0.01	0.20 ± 0.06
		Gen. HC	0.55 ± 0.10	0.09 ± 0.20	0.28 ± 0.13
		EDA-spec. HC	0.59 ± 0.08	0.16 ± 0.13	0.37 ± 0.13
		Mantis	0.64 ± 0.07	0.22 ± 0.11	0.40 ± 0.12
		MOMENT	0.65 ± 0.04	0.24 ± 0.07	0.42 ± 0.10
		Chronos	0.64 ± 0.04	0.23 ± 0.08	0.41 ± 0.10
		UME	0.63 ± 0.09	0.20 ± 0.13	0.38 ± 0.11
Dreamt	Sleep/Wake	Dummy	0.50 ± 0.01	-0.01 ± 0.01	0.58 ± 0.04
		Gen. HC	0.65 ± 0.04	0.32 ± 0.11	0.69 ± 0.06
		EDA-spec. HC	0.69 ± 0.01	0.38 ± 0.03	0.72 ± 0.02
		Mantis	0.73 ± 0.02	0.46 ± 0.03	0.76 ± 0.02
		MOMENT	0.69 ± 0.01	0.39 ± 0.02	0.72 ± 0.01
		Chronos	0.73 ± 0.01	0.47 ± 0.02	0.76 ± 0.01
		UME	0.70 ± 0.02	0.41 ± 0.03	0.74 ± 0.02
HHISS	Stress/calm	Dummy	0.50 ± 0.00	-0.00 ± 0.00	0.23 ± 0.09
		Gen. HC	0.45 ± 0.04	-0.10 ± 0.07	0.34 ± 0.04
		EDA-spec. HC	0.55 ± 0.05	0.09 ± 0.11	0.45 ± 0.07
		Mantis	0.51 ± 0.02	0.20 ± 0.03	0.50 ± 0.03
		MOMENT	OoM	OoM	OoM
		Chronos	OoM	OoM	OoM
		UME	0.49 ± 0.06	-0.02 ± 0.12	0.38 ± 0.07
HeartS	Sleep/Wake	Dummy	0.49 ± 0.00	-0.01 ± 0.01	0.08 ± 0.02
		Gen. HC	0.71 ± 0.06	0.35 ± 0.11	0.39 ± 0.08
		EDA-spec. HC	0.72 ± 0.14	0.36 ± 0.25	0.41 ± 0.20
		Mantis	0.75 ± 0.16	0.40 ± 0.28	0.45 ± 0.23
		MOMENT	0.69 ± 0.12	0.24 ± 0.16	0.31 ± 0.12
		Chronos	0.74 ± 0.13	0.35 ± 0.21	0.40 ± 0.19
		UME	0.73 ± 0.17	0.32 ± 0.25	0.38 ± 0.18
WESAD	Low/High Arousal	Dummy	0.55 ± 0.05	0.08 ± 0.07	0.15 ± 0.07
		Gen. HC	0.56 ± 0.15	0.08 ± 0.22	0.26 ± 0.11
		EDA-spec. HC	0.59 ± 0.20	0.08 ± 0.31	0.32 ± 0.14
		Mantis	0.58 ± 0.11	0.13 ± 0.19	0.32 ± 0.13
		MOMENT	OoM	OoM	OoM
		Chronos	0.56 ± 0.18	0.07 ± 0.28	0.28 ± 0.20
		UME	0.56 ± 0.12	0.08 ± 0.18	0.28 ± 0.15
WESAD	Low/High Valence	Dummy	0.51 ± 0.03	0.01 ± 0.04	0.66 ± 0.06
		Gen. HC	0.54 ± 0.20	0.08 ± 0.33	0.76 ± 0.12
		EDA-spec. HC	0.54 ± 0.19	0.04 ± 0.30	0.62 ± 0.22
		Mantis	0.61 ± 0.08	0.18 ± 0.13	0.71 ± 0.13
		MOMENT	OoM	OoM	OoM
		Chronos	0.45 ± 0.12	-0.10 ± 0.20	0.59 ± 0.21
		UME	0.63 ± 0.13	0.18 ± 0.18	0.73 ± 0.13
APSYNC	Low/High engagement	Dummy	0.45 ± 0.20	0.13 ± 0.13	0.13 ± 0.13
		Gen. HC	0.81 ± 0.20	0.32 ± 0.32	0.72 ± 0.29
		EDA-spec. HC	0.81 ± 0.20	0.32 ± 0.32	0.72 ± 0.29
		Mantis	0.86 ± 0.15	0.41 ± 0.43	0.82 ± 0.19
		MOMENT	0.48 ± 0.32	-0.05 ± 0.10	0.33 ± 0.30
		Chronos	0.61 ± 0.40	0.18 ± 0.86	0.70 ± 0.32
		UME	0.89 ± 0.22	0.49 ± 0.57	0.83 ± 0.33