Benchmarking Foundation Models on Out-of-Distribution Wearable Biosignals

Andres Alvarez Olmo[1], Sotirios Vavaroutas[1], Yu Wu[1], Cecilia Mascolo[1]

***Abstract*— Foundation models (FMs), including large language models and time series-specific models, have shown promise in the mobile and wearable data domain, particularly for analysing ECG and EEG biosignals. However, their performance on out-of-distribution (OOD) time series, especially when collected from diverse users and devices, remains underexplored. While task-specific models excel in in-domain tasks, they often struggle with generalisation, particularly when data quality varies across different mobile and wearable devices. This work evaluates the robustness of FMs on OOD time series from diverse sensor technologies, highlighting their strengths and limitations in mobile and wearable real-world applications.**

I. Introduction

Biosignal time series play a key role in the monitoring and diagnosis of physiological conditions [1]. Integrating biosignals such as electrocardiogram (ECG) and electroencephalogram (EEG) into mobile and wearable health technologies improves the accessibility of medical care worldwide [2]. However, biosignals differ from generic time series data due to their non-stationarity and complex temporal dynamics [3]. Foundation models have emerged as a promising approach to improve generalisation across diverse sensor technologies and recording settings as they do not require large annotated datasets [4]. Importantly, they offer advantages over task-specific models, which are often designed in an end-to-end fashion for in-domain tasks [5]. Despite their potential, there are limited studies benchmarking them on biosignal data, particularly in out-of-distribution settings. OOD data encompasses: dataset shifts, where variations occur due to sensors or collection differences within the same domain; and domain shifts, where the underlying data distribution changes across different contexts or applications. We seek to address this gap by assessing the robustness of FMs in real-world mobile and wearable health applications.

II. Methods

Foundation Models We test diverse FMs from a range of real-world use cases. The first FM explored is a transformer-based model, namely ECG-FM, which is pretrained on 2.5 million samples using ECG augmentations and contrastive learning [6]. Further, we examine WildECG, a state-space model for representation learning pretrained on 75,000 ECG recordings [7], and EEGNetv4, a compact convolutional neural network used for EEG interpretation [8].

Datasets Tested We examine annotated ECG datasets, including clinical (12-lead) and non-clinical (3-lead) variants, as well as annotated EEG datasets from real-world settings.

III. Evaluation

Case Studies We evaluate the predictive performance of selected foundation models using accuracy, F1-score, sensitivity, specificity and precision on OOD data. Specifically, we compare model performance before fine-tuning (zero-shot setting) and after different fine-tuning strategies. This approach enables us to assess their domain adaptation capabilities across varying dataset conditions.

[1]University of Cambridge

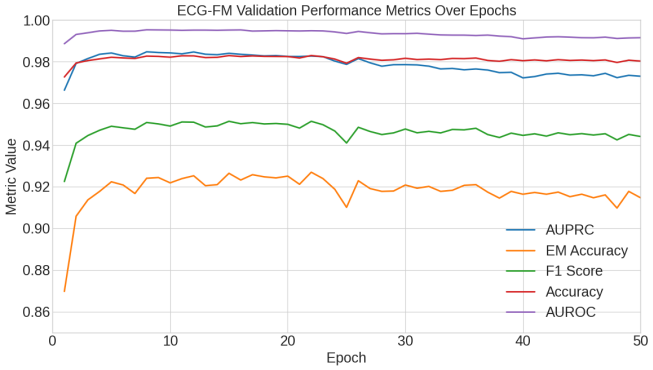

Fig. 1. ECG-FM validation metrics during fine-tuning on a class-balanced version of CODE-15, reaching convergence at epoch 22.

Results Performance metrics reveal a clear gap between zero-shot and fine-tuned models. While raw pre-trained models fail to solve downstream tasks effectively, applying minimal fine-tuning yields highly promising results. For ECG, accuracy differs by up to 50% in stress detection tasks based on 3-lead non-clinical data, and by up to 40% in heart condition classification tasks using 12-lead clinical data. Similarly, for EEG, we observe accuracy differences of up to 35%. These substantial gains result from light fine-tuning, as illustrated in Figure 1, where models reach convergence within the initial epochs. Although knowledge acquired during pre-training proves transferable, foundation models on their own are not yet sufficient to generalise across OOD datasets recorded using different sensor configurations.

IV. Conclusion

This study examines the generalisation capabilities of ECG and EEG foundation models on out-of-distribution data. Our analysis reveals that although pre-trained models perform poorly in zero-shot settings, minimal fine-tuning significantly improves their adaptability across datasets captured using diverse sensor technologies and varying electrode configurations. These findings demonstrate the potential of FMs for practical biosignal analysis, particularly in mobile and wearable healthcare systems facing distributional variability.

References

[1] D. Yoon, J.-H. Jang, B. Choi, and C. Han, "Discovering hidden information in biosignals from patients by artificial intelligence," 2020.
[2] A. Subasi, "Biomedical signal analysis and its usage in healthcare," 2019.
[3] A. Karagiannis, P. Constantinou, and D. Vouyioukas, "Biomedical time series processing and analysis methods: The case of empirical mode decomposition," 2011.
[4] Y. Han and C. Ding, "Foundation models in electrocardiogram: A review," 2024.
[5] R. Bommasani, D. A. Hudson, E. Adeli, Altman *et al.*, "On the opportunities and risks of foundation models," 2021.
[6] K. McKeen, L. Oliva, S. Masood, A. Toma, B. Rubin, and B. Wang, "ECG-FM: An open electrocardiogram foundation model," 2024.
[7] K. Avramidis, D. Kunc, B. Perz, Adsul *et al.*, "Scaling representation learning from ubiquitous ECG with state-space models," 2024.
[8] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces," 2018.