December 2024

An Electrocardiogram Multi-task Benchmark with Comprehensive Evaluations and Insightful Findings

Yuhao XU [a] Jiaying LU [c] Sirui DING [b] Defu CAO [d] Xiao HU [c] and Carl YANG [a,1]
[a] *Department of Computer Science, Emory University*
[b] *Bakar Computational Health Sciences Institute, University of California, San Francisco*
[c] *Center for Data Science, School of Nursing, Emory University*
[d] *Department of Computer Science, University of Southern California*

Abstract. In the process of patient diagnosis, non-invasive measurements are widely used due to their low risks and quick results. Electrocardiogram (ECG), as a non-invasive method to collect heart activities, is used to diagnose cardiac conditions. Analyzing the ECG typically requires domain expertise, which is a roadblock to applying artificial intelligence (AI) for healthcare. Through advances in self-supervised learning and foundation models, AI systems can now acquire and leverage domain knowledge without relying solely on human expertise. However, there is a lack of comprehensive analyses over the foundation models' performance on ECG. This study aims to answer the research question: *"Are Foundation Models Useful for ECG Analysis?"* To address it, we evaluate language / general time-series / ECG foundation models in comparison with time-series deep learning models. The experimental results show that general time-series / ECG foundation models achieve a top performance rate of 80%, indicating their effectiveness in ECG analysis. In-depth analyses and insights are provided along with comprehensive experimental results. This study highlights the limitations and potential of foundation models in advancing physiological waveform analysis. The data and code for this benchmark are publicly available at `https://github.com/yuhaoxu99/ECGMultitasks-Benchmark`.

Keywords. Electrocardiogram Analysis, Foundation Models, Machine Learning

1. Introduction

Electrocardiogram (ECG) records the heart's electrical activities via skin-placed electrodes [1], producing waveforms that decipher cardiac functions. Its non-invasive nature and ease of collection make ECG ideal for continuous monitoring and early detection of cardiovascular abnormalities. ECG is used for diagnosing arrhythmias [2], myocardial infarctions [3] and analyzing heart rate variability [4], highlighting its diverse utilities. However, ECG analysis is challenging due to individual variations, complex waveforms, and susceptibility to noises [5]. Traditional ECG analysis relies on specialized clinicians, which is resource-intensive and does not scale well with large data volumes, increasing

[1] Corresponding Author: Carl Yang, j.carlyang@emory.edu.

the risk of diagnostic errors. Advances in artificial intelligence (AI) have led to AI-assisted ECG diagnostics surpassing human performance [6].

AI models enhance ECG analysis by extracting rich features. Beyond detecting heart diseases, ECG can infer age [7], gender [7], blood pressure [8], and potassium levels [9]. Al-Zaiti et al. [6] used a random forest model that outperformed clinicians and FDA-approved systems in detecting acute myocardial ischemia. While traditional machine learning relies on feature engineering, potentially losing clinically relevant information, neural networks can use raw ECG signals, preserving critical information. Baloglu et al. [10] achieved over 99% accuracy in myocardial infarction detection using a convolutional neural network. However, neural networks require extensive labeled data, which may not always be available. Foundation models address this by leveraging large-scale pre-training and task-specific fine-tuning. McKeen et al. [11] proposed ECG-FM, a transformer-based model pre-trained on 2.5 million samples, demonstrating the strong potential of unsupervised foundation models. Despite the emergence of ECG foundation models, fair and comprehensive evaluations on their effectiveness are lacking.

In this study, we construct a benchmark to fairly evaluate existing foundation models for ECG analysis, including large language model (LLM), time-series foundation model (TSFM), and an ECG foundation model (ECGFM), in contrast to traditional time-series deep learning model (TSDL). We compare their performance across five tasks, assessing ECG data modeling from different perspectives: simple feature extraction (RR interval estimation), complex feature extraction (age estimation), balanced labels (gender classification), imbalanced labels (potassium abnormality prediction), and multi-class classification (arrhythmia detection). Our evaluation scenarios encompass zero-shot, few-shot, and fine-tuning approaches. Through these comparisons, we analyze the strengths and weaknesses of different models and explore the effectiveness of foundation models. We envision our findings can inspire advancements in using foundation models for physiological waveform analysis. Our code is open-sourced to support future research.

2. Methods

Our experiment is conducted on the MIMIC-IV-ECG [12] dataset, which is currently the largest publicly accessible ECG dataset, comprising 800,035 diagnostic electrocardiograms from 161,352 unique patients. Each ECG strip is 12-lead and 10 seconds in length with 500 Hz sampling rate, denoted by $x \in \mathbb{R}^{C \times L}$ where $C = 12$ and $L = 10 \times 500 = 5000$.

Downstream tasks. We evaluate the performance of the benchmark on the following tasks: (1) ***RR Interval Estimation.*** The RR interval, which represents the time between two R-wave peaks in an ECG, is directly calculated from the ECG signal. (2) ***Age Estimation.*** Patient age estimation involves analyzing ECG signal characteristics to estimate age, challenging the model to effectively interpret complex signal patterns correlated with physiological aging. (3) ***Gender Classification.*** Gender classification is a binary classification task with a roughly balanced ratio of 50% to 50%. (4) ***Potassium Abnormality Prediction.*** We use ECG strips to predict the Potassium (blood) lab test result which is taken between ECG recording time and one hour after the ECG time. This task is challegning, with imbalanced ratio of 97% (normal) to 3% (abnormal). (5) ***Arrhythmia Detection.*** We select the 14 most frequently occurring diagnoses, with the

remaining ones grouped under "Others", resulting in a total of 15 labels. Among these downstream tasks, RR interval estimation and age estimation are regression tasks, where the prediction target $y \in \mathbb{R}$. Gender prediction and potassium abnormality prediction are binary classification tasks, where the prediction target $y \in \{0,1\}$. Arrhythmia detection is multiclass classification task, where the prediction target $y \in \{1,2,3,\ldots,M\}$ ($M = 15$ denotes the phenotype of arrhythmia).

Evaluated Models. We select the following models for benchmarking: TimesNet [13], DLinear [14], GPT-2 [15], Llama 3.1 [16], MOMENT [17], TEMPO [18] and ECG-FM [11]. The details of these models are shown in Table 1. For the TSDL, TSFM, and ECGFM model categories, the original data are downsampled to match the input length required by the pre-trained models. For the LLM, the input is designed as a prompt based on features calculated from the original ECG data.

Table 1. Pre-Trained Datasets and Tasks of Benchmark Models. For datasets, only UCR/UEA, TSB-UAD, PhysioNet 2021, MIMIC-IV-ECG, UNH-ECG contains ECG data. For tasks, both waveform forecast and autoregressive use observed data to predict future time steps, masked time-series prediction can involve predict time-series masked out in the middle, while ECG-FM propose variants of masking tasks tailed for ECG.

Category	Models	Pre-Trained Dataset	Pre-Train Tasks
TSDL	TimesNet	ETT, Monash	waveform forecast
	DLinear	ETT, Monash	waveform forecast
LLM	GPT2	WebText	autoregressive
	Llama3.1	Meta internal corpus	autoregressive
TSFM	MOMENT	ETT, Monash, UCR/UEA, TSB-UAD	masked time-series prediction
	TEMPO	ETT, Monash	waveform forecast
ECGFM	ECG-FM	PhysioNet 2021, MIMIC-IV-ECG, UHN-ECG	wav2vec 2.0 Masking, CMSC, RLM

3. Results

Table 2. Benchmarking experimental results. Highlighted are the top first, second, and third results. (RR Interval Estimation, Age Estimation, Gender Classification, Potassium Abnormality Prediction, Arrhythmia Detection, and zero-shot, few-shot, fine-tune are denoted as RR., Age, Gen., Ka, AD, and zs, fs, ft respectively.)

			TimesNet	DLinear	GPT2	LLama3.1	MOMENT	TEMPO	ECG-FM
Regre. (MAE)↓	RR.	zs	817.0 ± 2.5	816.4 ± 2.9	816.0 ± 2.5	815.1 ± 1.9	816.6 ± 2.1	816.3 ± 1.9	816.3 ± 1.9
		fs	814.9 ± 1.9	816.0 ± 2.5	816.2 ± 1.2	816.2 ± 1.2	801.0 ± 1.4	808.2 ± 2.6	698.2 ± 96.7
		ft	304.3 ± 4.3	786.0 ± 5.4	823.1 ± 5.8	822.3 ± 3.1	146.9 ± 1.3	141.5 ± 2.1	147.3 ± 1.3
	Age	zs	62.28 ± 0.36	62.63 ± 0.50	62.40 ± 0.38	62.66 ± 0.14	62.61 ± 0.37	62.33 ± 0.38	62.27 ± 0.38
		fs	62.58 ± 0.38	61.87 ± 0.32	62.30 ± 0.19	62.30 ± 0.19	46.95 ± 2.16	54.79 ± 1.64	19.58 ± 5.95
		ft	24.89 ± 0.07	28.46 ± 0.74	61.86 ± 0.56	61.61 ± 0.72	13.41 ± 0.45	13.52 ± 0.31	13.49 ± 0.17
Binary Class (F1 Score)↑	Gen.	zs	0.60 ± 0.00	0.51 ± 0.08	0.00 ± 0.00	0.20 ± 0.00	0.34 ± 0.00	0.42 ± 0.00	0.33 ± 0.00
		fs	0.33 ± 0.00	0.49 ± 0.14	0.01 ± 0.00	0.04 ± 0.00	0.52 ± 0.03	0.36 ± 0.03	0.33 ± 0.00
		ft	0.51 ± 0.05	0.57 ± 0.01	0.04 ± 0.00	0.06 ± 0.00	0.68 ± 0.02	0.53 ± 0.01	0.34 ± 0.02
	Ka	zs	0.06 ± 0.00	0.05 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.18 ± 0.00	0.35 ± 0.00
		fs	0.00 ± 0.00	0.06 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.49 ± 0.00	0.49 ± 0.00	0.34 ± 0.22
		ft	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.00	0.02 ± 0.00	0.49 ± 0.00	0.50 ± 0.00	0.49 ± 0.00
15 Class (ACC)↑	AD	zs	0.06 ± 0.01	0.03 ± 0.02	0.04 ± 0.00	0.48 ± 0.01	0.03 ± 0.02	0.02 ± 0.00	0.07 ± 0.00
		fs	0.01 ± 0.00	0.40 ± 0.01	0.48 ± 0.00	0.48 ± 0.00	0.49 ± 0.01	0.43 ± 0.07	0.18 ± 0.21
		ft	0.03 ± 0.00	0.48 ± 0.02	0.49 ± 0.02	0.46 ± 0.06	0.66 ± 0.03	0.54 ± 0.14	0.49 ± 0.03
Benchmark (Win Rate)↑	Overall	zs	20%	0%	0%	40%	0%	0%	40%
		fs	0%	0%	0%	0%	60%	20%	20%
		ft	0%	0%	0%	0%	60%	40%	0%
		ALL	6.7%	0%	0%	13.3%	40%	20%	20%

As the performance and interpretation results presented in Table 2 and Figure 1, respectively, the following observations can be made from the experimental results.

① **Direct application of LLM on ECG is infeasible, showing inferior performance compared to TSDL.** LLMs struggled to determine gender and age from ECG, may due to insufficient knowledge linking demographic information to ECG features.

② **Models pre-trained on time-series or ECG data outperform LLMs significantly.** LLMs are better suited for text tasks than time-series data processing. Extracting features for prompt design may lead to loss of crucial temporal information.

③ **Specialized ECGFM did not significantly outperform TSFM.** TSFM's pre-training likely provides a robust understanding of time-series dynamics, enabling good adaptation to various tasks, compensating for the lack of ECG-specific training.

④ **Foundation model requires sufficient fine-tuning samples, as zero and few-shot performance was not good enough.** Differences between ECG data and pretraining data mean limited tuning may hinder effective understanding of ECG-specific tasks.

⑤ **Foundation model provides more interpretable results than TSDL.** Figure 1 shows the saliency maps for RR interval estimation. TSFM and ECG-FM effectively capture the feature peaks, demonstrating greater interpretability.

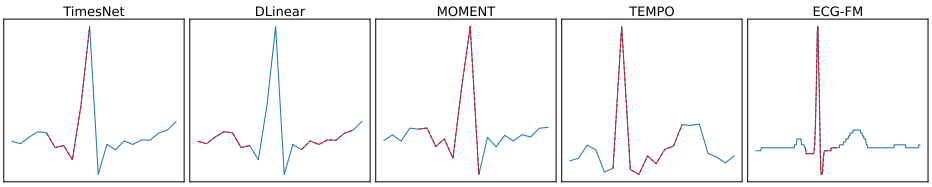

Figure 1. Saliency maps for the ECG-based RR Interval Estimation task. The blue line represents the ECG signal, the red line highlights the features the model focuses on. The same ECG segment is used, with downsampling applied due to varying input lengths of the pretrained models.

4. Discussion

From the results presented in Table 2, we can conclude that time-series foundation models are effective for ECG analysis. Based on the five summarized insights above, we discuss how these insights help us develop more advanced ECG foundation models in the future. From ①, applying LLM to ECG still needs specialized prompt design and external knowledge, which might require a retrieval-augmented generation (RAG) technique. Point ② addresses the importance of pre-training large-scale foundation models for various ECG downstream tasks. Moreover, based on ③, the pre-training data may include more than just ECG. The inclusion of general domain time-series could also boost the model performance on ECG. The current state-of-the-art TSFM and ECGFM still need amounts of fine-tuning samples as stated in ④. This motivates more efforts in the future to develop more advanced methods to pre-train or adapt the foundation model on ECG under zero and few-shots settings. From ⑤, the foundation model provides better interpretability compared to TSDL which paves a new way to explainable AI (XAI) in ECG analysis.

5. Conclusion

In this study, we build a comprehensive benchmark to evaluate various deep learning and foundation models for ECG analysis. Our results indicate that while time-series foundation models and ECG foundation models exhibit strong performance in certain tasks, suggesting their usefulness for ECG analysis, large language models struggle with ECG data, emphasizing the need for domain-specific and task-specific pre-training. Overall, our findings highlight the strengths and limitations of different foundation models for ECG analysis, underscoring the importance of foundation models and robust benchmarks for them.

6. Acknowledgement

This research was partially supported by the US National Science Foundation under Award Number 2319449 and Award Number 2312502, as well as the US National Institute of Diabetes and Digestive and Kidney Diseases of the US National Institutes of Health under Award Number K25DK135913.

References

[1] Majd AlGhatrif and Joseph Lindsay. A brief review: history to understand fundamentals of electrocardiography. *Journal of community hospital internal medicine perspectives*, 2012.

[2] Ary L Goldberger and et al. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 2000.

[3] U Rajendra Acharya and et al. Application of deep convolutional neural network for automated detection of myocardial infarction using ecg signals. *Information Sciences*, 2017.

[4] Jakub Parák and Jan Havlík. Ecg signal processing and heart rate frequency detection methods. *Proceedings of Technical Computing Prague*, 2011.

[5] U Rajendra Acharya and et al. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, 2017.

[6] Salah S Al-Zaiti and et al. Machine learning for ecg diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, 2023.

[7] Zachi I Attia and et al. Age and sex estimation using artificial intelligence from standard 12-lead ecgs. *Circulation: Arrhythmia and Electrophysiology*, 2019.

[8] Monika Simjanoska and et al. Non-invasive blood pressure estimation from ecg using machine learning techniques. *Sensors*, 2018.

[9] Philipp von Bachmann and et al. Evaluating regression and probabilistic methods for ecg-based electrolyte prediction. *Scientific Reports*, 2024.

[10] Ulas Baran Baloglu and et al. Classification of myocardial infarction with multi-lead ecg signals and deep cnn. *Pattern recognition letters*, 2019.

[11] Kaden McKeen and et al. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024.

[12] Brian Gow and et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset.

[13] H Wu et al. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023.

[14] Ailing Zeng and et al. Are transformers effective for time series forecasting? In *AAAI*, 2023.

[15] Alec Radford and et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[16] Abhimanyu Dubey and et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] Mononito Goswami and et al. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[18] Defu Cao and et al. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948*, 2023.