Are Time Series Foundation Models Ready for Vital Sign Forecasting in Healthcare?

Xiao Gu* XIAO.GU@ENG.OX.AC.UK
Yu Liu* YU.LIU@ENG.OX.AC.UK
Zaineb Mohsin ZAINEB.MOHSIN@NDCN.OX.AC.UK
Jonathan Bedford JONATHAN.BEDFORD@NDCN.OX.AC.UK
Anshul Thakur ANSHUL.THAKUR@ENG.OX.AC.UK
Peter Watkinson PETER.WATKINSON@NDCN.OX.AC.UK
Lei Clifton LEI.CLIFTON@PHC.OX.AC.UK
Tingting Zhu TINGTING.ZHU@ENG.OX.AC.UK
David A. Clifton DAVID.CLIFTON@ENG.OX.AC.UK
University of Oxford, United Kingdom

Abstract

The rise of foundation models, particularly large language models like ChatGPT, has revolutionized natural language processing and demonstrated remarkable generalization across numerous healthcare applications. Building on this success, foundation models for time series forecasting have emerged, offering new opportunities by leveraging pretraining on large-scale datasets. However, existing time series foundation models are pretrained with minimal clinical data, and their potentials for continuously recorded clinical time series, such as vital signs, remain largely under-explored. This motivates our endeavor to integrate time series foundation models with vital sign data to address critical clinical challenges, particularly in predicting patient deterioration. Through an extensive evaluation of various settings and configurations of these models, alongside comparisons with conventional forecasting models, we highlight the significant opportunities for improvement in developing clinically useful time series forecasting models. In a word, the *"ChatGPT"* moment for time series foundation models, in the typical clinical domain, is yet to come.

Keywords: Time series forecasting, vital sign, clinical deterioration, foundation models

Data and Code Availability We use the following data: (1) the eICU collaborative research database (Pollard et al., 2018), (2) the MIMIC-III waveform database (Johnson et al., 2016), which are both available on the PhysioNet (Goldberger et al., 2000). Details for dataset processing and code implementation are provided in Appendix A and Appendix B, respectively.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Vital signs are the simplest, and probably the most important medical information collected from patients in varied phases of healthcare delivery, from hospitals to community settings (Kellett and Sebat, 2017). They act as the major components of existing established early warning score (EWS) systems, and recent advances have demonstrated significant capabilities in predicting adverse clinical events with these vital signs (Al-Shwaheen et al., 2022; Topol, 2024).

The acquisition of vital signs in clinical practice is undergoing a paradigm shift, transiting from infrequent, intermittent routine observations to continuous data streaming, enabled by sophisticated bedside monitors in intensive care units (ICU) and potentially wearable devices beyond ICU (Shamout et al., 2019; Gu et al., 2023). Such high-frequency, long-term vital-sign time series allow a holistic understanding of patient health trajectories (Harutyunyan et al., 2019), as well as provide actionable insights for timely recognition/intervention of patient deterioration (Clifton et al., 2012). Among these, one of the fundamental tasks, is to develop preventive solutions based on predictive modeling and forecasting

* These authors contributed equally
Corresponding: XG, YL, TZ, DAC

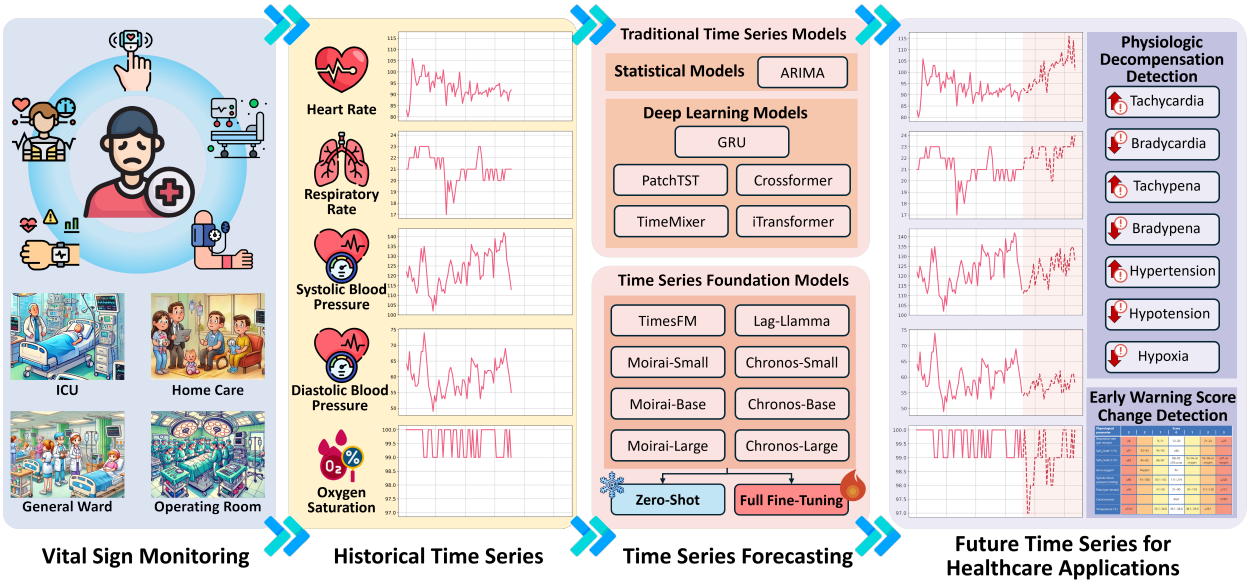


Figure 1: Illustration of our proposed pipeline integrating Time Series (Foundation) Models with vital sign forecasting, in healthcare settings. With the continuous acquisition of vital signs, we leverage the historical time series as input for the task of vital sign forecasting. Different time series forecasting models, including statistical methods, conventional deep learning models, and foundation models are applied. The forecasted vital sign values are, subsequently, investigated in relation to their predictive accuracy for clinical deterioration events.

of these vital-sign time series (Forkan et al., 2017; da Silva et al., 2021; Topol, 2024).

On the other hand, recent advances in deep learning, especially transformer architectures and self-supervised learning paradigms, are transforming the realm of time series analysis (Liang et al., 2024; Wang et al., 2024b). In particular, inspired by the remarkable success of large language models (LLMs) in natural language processing area, a series of time series foundation models (TSFMs) are emerging as powerful tools for time series forecasting (Das et al., 2023; Rasul et al., 2024; Woo et al., 2024; Ansari et al., 2024; Gao et al., 2024). These TSFMs have demonstrated superior forecasting performance in a variety of domains (e.g., traffic, weather, finance), even achieving satisfactory performance in a zero-shot manner. Such superiority positions them as promising candidates for healthcare applications in vital sign forecasting and patient outcome prediction.

In fact, most TSFMs are pretrained on minimal or even no clinical vital sign data (refer to Appendix Table A2). Despite their impressive performance in a wide range of non-clinical benchmarks (Das et al., 2023; Rasul et al., 2024; Woo et al., 2024; Ansari et al., 2024), their potential usage in clinical vital sign forecasting and associated performance measurement are not investigated yet. Clinical vital sign data, with its unique properties such as variability across patients and instability of patient conditions, presents novel challenges that may affect model performance (Afshar et al., 2021). There remains a lack of clear insights into what factors drive the effectiveness of TSFMs in clinical contexts and how these models can be optimized for real-world healthcare applications.

In this paper, we take a step forward, and aim to look into the following three key research questions concerning the application of TSFMs in clinical vital sign forecasting:

Model Development and Evaluation. *How can time series forecasting models be developed and evaluated in the context of clinical vital signs, and what are the key metrics for this evaluation?* By integrating time series forecasting approaches with continuous vital sign monitoring (Figure 1), we present a novel endeavor to systematically investigate its potential applications in healthcare, particularly for patient deterioration prediction.

Foundation *versus* Conventional Models. *How is the performance of TSFMs in clinical vital sign forecasting, compared to conventional time series*

forecasting models? We performed a series of experiments comparing a wide spectrum of methodologies, to gain insights into the effective utilization of these models (Section 4.1).

Attributing Factors of TSFMs. *What factors influence the performance of TSFMs and conventional models when applied to clinical vital sign data?* We looked into several critical issues, regarding duration of historical windows (Section 4.2), multi-/univariate settings (Section 4.3), computing efficiency (Section 4.6), generalization capability (Section 4.5), that are related to practical clinical use.

2. Related Work

Time Series Forecasting. Time series forecasting is a well-studied task, with considerable research efforts dedicated to model development and dataset/benchmark curation (Benidis et al., 2022). The advancements of deep learning have enabled the evolution of models from traditional statistical methods to deep architectures, allowing to automatically capture the complex temporal dependencies from raw data. Despite the success of these models in benchmarks across various non-clinical data domains, their practical application in the clinical domain, typically for vital sign forecasting, remains under-explored. There are, in fact, several fundamental differences that limits the direct generalization of these methodologies. Most time series forecasting pipelines are developed based on temporal data split, evaluated using straightforward quantitative measures, such as mean average errors. It remains open how to incorporate clinical perspective into such forecasting pipeline.

Time Series Foundation Models. Foundation models (FMs) typically benefit from pretraining on large-scale datasets in self-supervised learning manners. Their generalization capability across a wide spectrum of data domains and tasks, even via zero-shot learning, has been highlighted (Liang et al., 2024). Particularly, in the realm of time series, the development of foundation models has primarily focused on the generalization of forecasting capability across different data domains. Nevertheless, these models (Das et al., 2023; Rasul et al., 2024; Woo et al., 2024; Ansari et al., 2024; Gao et al., 2024) differ in several key aspects of their design, including whether they handle uni- or multivariate inputs, the length of context windows, and whether they incorporate probabilistic inference (refer to Appendix Table A2). These settings, along with their actual performance, are crucial in selecting appropriate FMs for clinical vital sign forecasting, an area that has yet to be systematically investigated.

Physiological Trajectory Predictive Modeling. Continuous modeling of individual physiological trajectories, such as vital signs, is important for early detection and prediction of potential deterioration, especially for critically ill patients. Several data analytic approaches, including unsupervised novelty detection (Clifton et al., 2012) and end-to-end deep learning (Forkan et al., 2017; Shamout et al., 2019), have been developed to predict impending clinical deteriorations. In general clinical practice, such as the National Early Warning Score version 2 (NEWS2) in the UK and the Modified Early Warning Score (MEWS) in the USA, we oversee changes in vital signs as key indicators of health conditions. Such EWS systems reflect physiological deterioration based on well-established clinical practices, and have been widely applied in medical care, especially in intensive care (Gerry et al., 2024; Henry et al., 2015). In this sense, by predicting vital sign changes in the future, vital sign forecasting offers the potential to timely predict upcoming physiological decompensations and beyond (Sundrani et al., 2023). However, although the integration of vital sign forecasting with clinical outcome prediction is conceptually elegant, the validation pipeline and potential performance of these models, incorporating clinical insights, have yet to be explored.

3. Benchmarks and Tasks

3.1. Task Settings

Typically, for a time series forecasting task, as illustrated in Figure 2, the objective is to predict future values over a horizon length of P, based on the input historical window, with a window length of H. In practice, we set H and P as 6 and 3 hours, following the suggestions from clinicians. The impact of H is further discussed in Section 4.2.

To assess the performance of the predicted time series over the future P-hour horizon, two groups of evaluation metrics are employed as below.

3.1.1. General time series metrics

In line with general time series forecasting settings, we leverage a series of quantitative measurements, including Mean Average Error (MAE) and

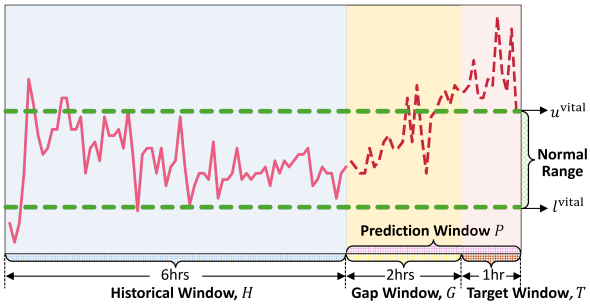


Figure 2: Illustration of our task settings.

Mean Squared Error (MSE). These have been widely adopted in benchmarking non-clinical data, measuring the numerical distances between forecasted values against ground truth.

3.1.2. Clinical deterioration metrics

Our further objective is to associate the predicted time series with relevant clinical outcomes. To ensure that the forecasted values are not only numerically accurate but also clinically meaningful, we go beyond the general time series forecasting metrics. We propose two additional validation protocols to assess the practical utility of the forecasted values, in clinical settings.

In this context, we introduce another variable, the target window T, as described in Figure 2. This is defined as the last 1 hours within the prediction window P. We aim to determine whether this target window, with a gap window of G between the historical window H, would effectively capture the changes of patient physiological trajectories, particularly those related to acute deteriorations. In practice, we introduce the following two subtasks to assess such clinical relevance, with further discussion available in Appendix Section C.

Physiological decompensation prediction. Aiming at the target window, we investigate whether the forecasted vital sign values can indicate the actual future deviations from normal physiological ranges. These deviations are associated with varied forms of physiological instability as listed in Table 1. In practice, we define such deviation as persist abnormality within the target window T. In this setting, we select samples whose historical windows do not present over-1 hour persistent abnormalities as normal-history samples, and assess whether these forecasted vital sign values can predict the normality/abnormality in the upcoming target window of these normal-history samples.

In detail, regarding the abnormality of vital signs, we defined those abnormalities as the persistent deviations outside the normal range, as outlined in Table 1. Let x_i^{vital} represent the values of the vital sign vital (vital $\in \{\text{HR}, \text{RR}, \text{SBP}, \text{DBP}, \text{SpO2}\}$) at time step i, then the abnormality target y is defined as below,

$$y = \mathbb{I}\left(\min_{i\in(0,T]} x_{H+G+i}^{\text{vital}} > u^{\text{vital}}\right) + \mathbb{I}\left(\max_{i\in(0,T]} x_{H+G+i}^{\text{vital}} < l^{\text{vital}}\right), \quad (1)$$

where l^{vital} and u^{vital} are the lower and upper bounds of the normal range of corresponding vital sign.

Table 1: Normal ranges of vital signs.

Vital Sign	Condition	Values
Heart Rate	Tachycardia	HR>100
	Bradycardia	HR<60
Blood Pressure	Hypertension	SBP>120 or DBP>80
	Hypotension	SBP<90 or DBP<60
Respiratory Rate	Tachypena	RR>18
	Bradypena	RR<12
SpO2	Hypoxia	SpO2<96

Early warning score change prediction. On the other hand, we leverage a validated composite measure for patient decompensation (Sundrani et al., 2023), NEWS2. By aggregating the corresponding scores associated with available vital sign values, we aim to investigate the efficacy of the forecasted values in actual detection of EWS changes. In practice, we focus on the average values of the last 1 hour of the historical window, as well as the last 1 hour of the prediction window (i.e. target window). The NEWS2 score is calculated based on the average vital sign values within the window of interest, considering only the available vital sign types. We define three types of outcomes based on comparisons with the last 1 hour of the historical window: *unchanged, elevated, reduced*. This evaluation protocol allows evaluating the capability of the forecasted values to reflect changes in the NEWS2 score.

By only considering HR, SBP, RR, and SpO2, the EWS score NEWS2 is defined as below,

$$f^{\text{EWS}}(\boldsymbol{x}_{t_a:t_b}) = \text{NEWS2}(\bar{\boldsymbol{x}}_{t_a:t_b}^{\text{HR}}, \bar{\boldsymbol{x}}_{t_a:t_b}^{\text{SBP}}, \bar{\boldsymbol{x}}_{t_a:t_b}^{\text{RR}}, \bar{\boldsymbol{x}}_{t_a:t_b}^{\text{SpO2}}), \quad (2)$$

where $\bar{\boldsymbol{x}}_{t_a:t_b}^{\text{vital}}$ is the mean value of time series of corresponding vital sign from t_a (exclusively) to t_b.

Accordingly, the target of early warning score change is defined as below,

$$y = \begin{cases} 0 & \text{if } f^{\text{EWS}}(\boldsymbol{x}_{t_0:t_1}) = f^{\text{EWS}}(\boldsymbol{x}_{t_2:t_3}), \\ 1 & \text{if } f^{\text{EWS}}(\boldsymbol{x}_{t_0:t_1}) > f^{\text{EWS}}(\boldsymbol{x}_{t_2:t_3}), \\ 2 & \text{if } f^{\text{EWS}}(\boldsymbol{x}_{t_0:t_1}) < f^{\text{EWS}}(\boldsymbol{x}_{t_2:t_3}), \end{cases} \tag{3}$$

where, in our task settings, $t_0 = H-1, t_1 = H$, and $t_2 = T+G, t_3 = T+P$, aiming to detect changes between the last 1 hour of the historical window and the last 1 hour of the prediction window.

3.2. Dataset Curation

Dataset preparation. We derived high-frequency multi-parameter vital-sign time series from two public open-source clinical databases, namely the eICU collaborative research database (Pollard et al., 2018) and the MIMIC-III waveform database (Johnson et al., 2016). They provide long-term continuous numeric recordings of multi-parameter vital signs. We focus on five typical types of vital signs including heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), respiratory rate (RR), and oxygen saturation (SpO2).

All recordings were resampled to the same frequency by calculating the median values every 5-minute interval. Subsequently, we extracted segments of length $(H+P)$ from the long-term recordings, using a sliding window step of 2 hours. Segments with more than 1 hour of consecutive missing data were filtered out, and the remaining data was imputed using forward filling, with median values applied where forward filling was not feasible. The details of the curated datasets are available in Table 2.

Table 2: Dataset statistics.

Dataset	#Patient	#Sample	Frequency
eICU	6,199	95,420	5 Minute
MIMIC-III	1,089	18,163	5 Minute

Data split. We further randomly split each dataset, subject-wise, into train/validation/test subsets by 6:2:2, without subject overlap. It should be noted that, our settings, differ from the generic temporal split setting, which divides each time series into training, validation, and test subsets in chronological order. We argue that such setting may not fully examine the practical utility of time series forecasting in healthcare, as the ability to generalize to unseen, novel patients is crucial.

3.3. Experimental Settings

We adopted four open-sourced TSFMs, including **TimesFM** (Das et al., 2023), **Chronos** (small/base/large) (Ansari et al., 2024), **Moirai** (small/base/large) (Woo et al., 2024), and **Lag-Llama** (Rasul et al., 2024), whose configurations are varied in multiple aspects, as listed in Table A2 of the Appendix.

In particular, these FMs take varied forms of input time series, either multivariate or univariate only. For a fair comparison, we adopted the univariate setting for all methods and reported their performance on each type of vital signs accordingly. Please refer to Section 4.3 for further discussion related to univariate and multivariate settings. In particular, for FMs, we tested two different types of implementations, **zero-shot** and **full fine-tuning**. For zero-shot learning, all four approaches were implemented with their officially pretrained models, whereas for full fine-tuning, we followed the official fine-tuning implementations of Chronos, Moirai, and Lag-Llama in this study.

On the other hand, to provide a deeper understanding of the performance of time series forecasting models, especially in comparison to traditional approaches, we developed a series of conventional models trained from scratch for benchmarking purposes. These include **ARIMA** (Box et al., 2015), **GRU** (Cho et al., 2014), **PatchTST** (Nie et al., 2023), **Crossformer** (Zhang and Yan, 2023), **iTransformer** (Liu et al., 2024), and **TimeMixer** (Wang et al., 2024a).

4. Results and Discussions

4.1. Quantitative performance comparison

Following the previous stated experiment settings, we reported the results of MAE (general metrics), Macro-F1 of vital sign related physiological decompensation prediction (clinical metrics), and Macro-F1 of EWS change prediction (clinical metrics) in Tables 3, 4 and 5. The results of other metrics are available in Tables A3, A4 and A5 of the Appendix.

• *Numeric performance between train-from-scratch conventional methods, zero-shot of foundation models, and full fine-tuning of foundation models.* In terms of MAE, it is demonstrated in Table 3 that the zero-shot performance of most TSFMs is not competitive with even the trained-from-scratch performance of most conventional methods.

On the other hand, we observed consistent performance improvements, in terms of MAE, follow-

Table 3: Vital sign forecasting performance measured with Mean Absolute Error (MAE). **Bold** indicates the best result, and Underline indicates the second best result in each group.

Setting	Methods	eICU					MIMIC-III				
		HR	SBP	DBP	RR	SpO2	HR	SBP	DBP	RR	SpO2
Training from Scratch	ARIMA	4.588	10.465	4.982	2.620	1.326	4.466	11.058	5.560	2.483	1.263
	GRU	**4.462**	10.010	**4.733**	**2.492**	**1.251**	**4.331**	**10.722**	**5.328**	**2.364**	**1.180**
	PatchTST	4.479	10.044	4.744	2.513	1.256	4.351	10.760	5.372	2.388	1.200
	Crossformer	4.482	**9.974**	4.737	2.513	1.289	4.382	10.744	**5.328**	2.390	1.215
	iTransformer	4.480	10.042	4.766	2.502	1.253	4.371	10.779	5.406	2.376	1.202
	TimeMixer	4.479	10.027	4.748	2.503	1.256	4.376	10.791	5.373	2.386	1.210
Zero-Shot	TimesFM	5.228	11.393	5.326	2.765	1.512	4.973	11.992	5.900	2.583	1.442
	Chronos-Small	**4.807**	**11.130**	**5.181**	**2.668**	1.374	**4.609**	**11.962**	**5.884**	2.555	1.354
	Chronos-Base	4.821	11.356	5.256	2.732	1.393	4.667	12.082	5.900	**2.542**	1.334
	Chronos-Large	4.857	11.244	5.244	2.729	1.372	4.621	12.040	5.887	2.551	1.332
	Moirai-Small	5.355	12.141	5.737	2.816	1.563	5.259	12.971	6.497	2.703	1.528
	Moirai-Base	4.880	11.310	5.296	2.672	**1.367**	4.753	12.109	6.038	2.550	**1.316**
	Moirai-Large	5.069	11.453	5.381	2.671	1.431	4.934	12.259	6.109	2.556	1.378
	Lag-Llama	7.365	13.022	6.125	2.879	2.848	6.877	13.425	6.609	2.715	2.740
Full Fine-Tuning	Chronos-Small	4.547	10.270	4.833	**2.509**	1.309	**4.424**	**10.933**	5.454	**2.400**	1.229
	Chronos-Base	**4.546**	10.265	**4.829**	**2.509**	1.304	**4.424**	11.009	5.446	2.411	**1.220**
	Chronos-Large	4.586	**10.234**	4.844	2.614	1.310	4.436	10.960	**5.443**	2.414	1.229
	Moirai-Small	4.958	10.716	5.079	2.606	1.382	4.681	11.531	5.783	2.487	1.264
	Moirai-Base	4.835	10.629	5.002	2.561	**1.292**	4.825	12.162	5.686	2.452	1.234
	Moirai-Large	4.985	10.601	5.024	2.607	1.358	4.784	11.678	5.744	2.446	1.229
	Lag-Llama	5.970	11.259	5.427	2.558	1.570	5.297	12.145	6.020	2.463	2.719

Table 4: Physiological decompensation detection performance measured with macro-F1. **Bold** indicates the best result, and Underline indicates the second best result in each group.

Setting	Methods	eICU					MIMIC-III				
		HR	SBP	DBP	RR	SpO2	HR	SBP	DBP	RR	SpO2
Training from Scratch	ARIMA	**0.638**	0.581	0.568	0.547	0.571	**0.620**	0.520	0.557	0.590	0.566
	GRU	0.621	0.602	0.609	0.555	**0.598**	0.581	0.560	0.569	**0.615**	**0.617**
	PatchTST	0.607	**0.607**	**0.623**	0.543	0.585	0.532	**0.595**	0.579	0.571	0.578
	Crossformer	0.617	0.593	0.614	**0.560**	0.564	0.588	0.535	0.566	0.609	0.607
	iTransformer	0.625	0.593	0.614	0.557	0.596	0.598	0.561	0.576	0.607	0.587
	TimeMixer	0.632	0.593	0.614	0.553	0.561	0.596	0.545	**0.580**	0.605	0.568
Zero-Shot	TimesFM	0.578	0.570	0.544	0.534	0.519	0.533	**0.596**	0.564	0.560	0.497
	Chronos-Small	**0.613**	**0.573**	0.604	0.551	0.512	0.59	0.552	0.570	0.563	0.538
	Chronos-Base	**0.613**	0.567	0.599	0.542	0.518	**0.635**	0.561	**0.573**	**0.565**	0.540
	Chronos-Large	0.604	0.563	**0.606**	0.548	0.506	0.620	0.580	0.550	**0.565**	0.512
	Moirai-Small	0.573	0.556	0.546	0.564	0.543	0.528	0.526	0.544	0.538	0.543
	Moirai-Base	0.568	0.553	0.584	**0.570**	**0.548**	0.562	0.550	0.558	0.544	**0.575**
	Moirai-Large	0.569	0.554	0.568	0.561	0.498	0.561	0.507	0.548	0.529	0.515
	Lag-Llama	0.491	0.494	0.510	0.517	0.499	0.493	0.483	0.504	0.516	0.492
Full Fine-Tuning	Chronos-Small	0.599	0.585	0.606	0.569	0.513	**0.585**	0.540	0.527	0.564	0.494
	Chronos-Base	0.604	0.574	**0.611**	0.570	0.516	0.580	0.547	0.546	**0.586**	0.496
	Chronos-Large	**0.630**	**0.586**	**0.611**	0.555	0.516	0.565	**0.570**	**0.561**	0.574	0.505
	Moirai-Small	0.583	0.537	0.556	**0.577**	0.542	0.543	0.509	0.524	0.557	0.507
	Moirai-Base	0.516	0.532	0.579	0.563	**0.558**	0.533	0.496	0.531	0.529	**0.514**
	Moirai-Large	0.535	0.545	0.522	0.553	0.512	0.518	0.497	0.536	0.535	0.513
	Lag-Llama	0.532	0.577	0.557	0.544	0.504	0.553	0.497	0.556	0.555	0.192

ing the full fine-tuning of all TSFMs, although they still slightly fall behind the conventional methods, as shown in Table 3. This suggests that TSFMs pretrained on non-clinical datasets may struggle to master the transfer capability to clinical datasets, likely due to the inherent differences in time series characteristics across these domains.

• *General time series metrics versus clinical deterioration related metrics.* Despite the strong numerical accuracy of some models in general time series metrics, these models performed poorly in detecting abnormal clinical values as indicated in Table 4, highlighting a gap between traditional forecasting metrics and clinical relevance. This discrepancy is likely due to the fact that we only selected health-history samples for evaluation, to mimic the real-world scenarios where predicting abnormality for currently stable patients is crucial. However, the models were pretrained on the entire dataset, which also includes non-health-history samples. This suggests that selecting appropriate training samples for specific tasks could be crucial, and it is worth further exploring how tailored training sets might improve task-specific performance.

For the second sub-task, we derived the EWS for the last 1 hour within the target window for all testing samples and subsequently assessed whether the forecasted vital sign values could predict such changes. This approach provides a more comprehensive evaluation, considering all five vital signs as a composite measure. As shown in Table 5, the zero-shot performance of most TSFMs is competitive with conventional methods trained from scratch. In some cases, such as with TimesFM on F1, the TSFMs even outperform the others.

On the other hand, for some models, like Lag-Llama, we did not observe consistent and significant performance improvements after full fine-tuning. This may suggest that certain fine-tuning strategies, particularly those based on quantitative loss measurements like MSE, may not be effective in guiding the model to learn capabilities relevant to clinical deterioration. This points out a potential need for task-specific fine-tuning approaches that better align with clinical objectives.

4.2. Influence of historical window length

We also investigated the effects of the historical window length H on the prediction results. It is important to look into such parameter, since it indicates whether longer historical windows can empower the model with better capability to capture relevant patterns for future trend prediction, particularly for clinical vital signs. We show the results of HR and SBP in Figure 3, with comprehensive results listed in Figures A2 and A3 of the Appendix. In general, longer historical windows tend to improve performance in most cases. This may result in the need for longer recording periods, for more accurate vital sign forecasting, in clinical practice.

4.3. Univariate *versus* Multivariate

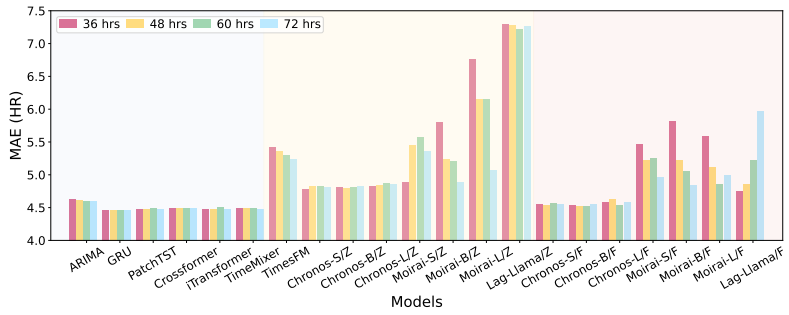
Vital signs are typically collected as multi-parameter forms. As mentioned in Section 3.3, due to the limitations of current FM configurations, and to ensure a fair comparison across different models, we conducted the experiments in univariate mode only. However, it would be beneficial to discuss whether incorporating additional vital signs types would result in improved performance of the forecasting of single vital sign. Such insights would be beneficial, to guide the development of future FMs for clinical applications.

We selected Moirai, which supports multivariate format as input, and reported its performance for both zero-shot and full fine-tuning under multivariate settings. We also reported all the results of conventional time series models with multiple vital signs as input. The results of HR and SBP forecasting are presented in Figure 4, with complete results available in Figures A4 and A5 of the Appendix.

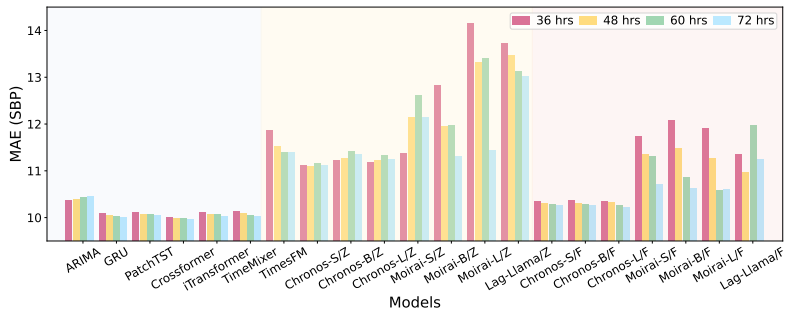
As shown in Figure 4, the performance of the TSFM Moirai, improved significantly when all five vital signs were incorporated for forecasting. However, a similar improvement was not observed in conventional time series models. This is a promising finding, suggesting that while FMs may underperform compared to conventional models in univariate mode, they hold potential in multivariate mode. This is likely due to the limited shared knowledge between non-clinical and clinical domains in terms of temporal characteristics, whereas the relationships between different variables in multivariate forms may be shared across domains.

4.4. Generalization across datasets

We further explored the generalization capability of GRU and Chronos-Small, across datasets and with the same vital sign type. The results, presented in Table 6, indicate that both models exhibit relatively good and similar performance in terms of generalization across dataset sources. This demonstrates the feasibility of pretraining a large clinical vital sign FM

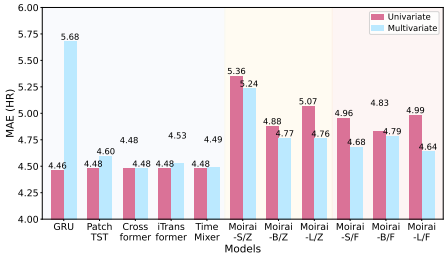


(a) HR

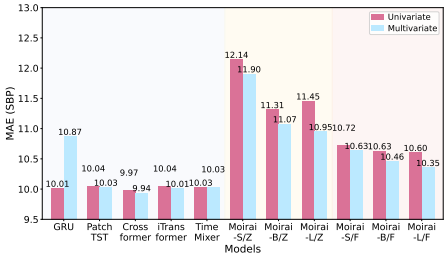


(b) SBP

Figure 3: Performance comparison of models with different historical lengths for vital sign forecasting on eICU dataset. "/Z" and "/F" denote zero-shot and full fine-tuning settings.



(a) HR



(b) SBP

Figure 4: Performance comparison of models trained in univariate and multivariate settings for vital sign forecasting on eICU dataset. "/Z" and "/F" denote zero-shot and full fine-tuning settings.

Table 5: Early warning score change detection performance comparison. **Bold** indicates the best result, and Underline indicates the second best result in each group.

Methods	eICU			MIMIC-III		
	F1	Rec.	Prec.	F1	Rec.	Prec.
ARIMA	**0.519**	0.514	0.550	0.531	0.523	0.556
GRU	0.501	0.500	0.562	0.526	0.517	0.580
PatchTST	0.506	0.504	0.564	0.522	0.514	0.578
Crossformer	**0.519**	**0.517**	**0.574**	**0.541**	**0.532**	**0.588**
iTransformer	0.498	0.497	0.563	0.511	0.502	0.570
TimeMixer	0.497	0.496	0.562	0.519	0.510	0.581
TimesFM	**0.532**	**0.529**	**0.539**	**0.530**	**0.527**	0.536
Chronos-Small	0.510	0.506	0.535	0.513	0.506	0.534
Chronos-Base	0.507	0.502	0.522	0.516	0.509	0.534
Chronos-Large	0.513	0.508	0.529	0.506	0.499	0.529
Moirai-Small	0.497	0.493	0.524	0.502	0.495	0.530
Moirai-Base	0.480	0.480	0.529	0.488	0.482	0.532
Moirai-Large	0.494	0.492	0.535	0.500	0.493	**0.542**
Lag-Llama	0.517	0.514	0.530	**0.530**	0.524	**0.542**
Chronos-Small	0.514	0.510	0.561	0.524	0.516	0.572
Chronos-Base	0.513	0.510	0.563	0.514	0.507	0.564
Chronos-Large	0.516	0.513	0.564	0.520	0.512	0.566
Moirai-Small	0.503	0.499	0.544	0.503	0.496	0.549
Moirai-Base	0.518	0.514	0.565	**0.525**	**0.517**	**0.572**
Moirai-Large	**0.529**	**0.524**	**0.568**	0.518	0.511	0.563
Lag-Llama	0.527	0.523	0.535	0.499	0.496	0.512

Table 6: Vital sign forecasting performance comparison (MAE) on eICU dataset. Chronos denotes Chronos-Small model. "@MIMIC" and "@eICU" denote the corresponding model trained on MIMIC-III and eICU, respectively.

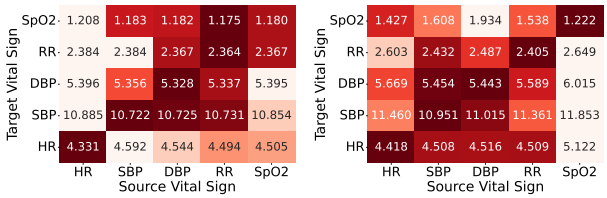
Model	HR	SBP	DBP	RR	SpO2
GRU@eICU	4.462	10.010	4.733	2.492	1.251
GRU@MIMIC	4.510	10.106	4.754	2.496	1.261
Chronos@eICU	4.547	10.270	4.833	2.509	1.309
Chronos@MIMIC	4.624	10.293	4.875	2.559	1.317

model and its potential for zero-shot performance on new clinical datasets.

4.5. Generalization across vital sign types

Furthermore, we aimed to explore the generalization capability of these time series forecasting models across different vital signs. Specifically, we assessed how well a model performs on one vital sign when it has been pretrained on a different vital sign type.

We compared the results of a conventional time series model GRU and one TSFM Chronos-Small,



(a) GRU (b) Chronos-Small/F

Figure 5: Cross vital sign forecasting performance comparison on MIMIC-III dataset. The color map is calculated by min-max normalization across each row. The annotated value in each block is MAE.

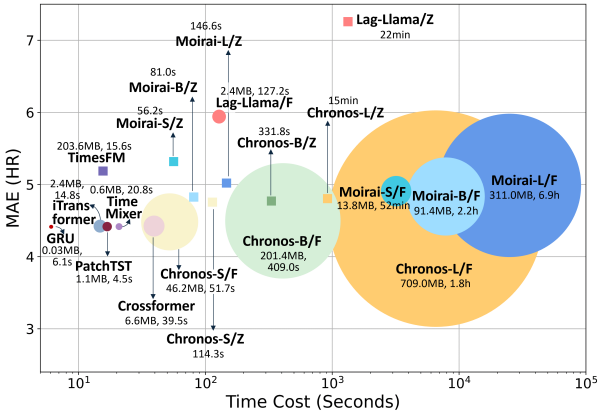


Figure 6: Computation efficiency, parameters and MAE of all models on eICU dataset. For conventional models and full fine-tuning (/F) models, we report the training time cost per epoch. For zero-shot (/Z) models, we report the inference time on the testing set. Note that the inference time for conventional forecasting models is less than 5 seconds. Marker size indicates the number of fine-tuned parameters.

with results presented in Figure 5. The generalization capability of the smaller GRU models seemed to perform better than the larger Chronos-Small. They may also suggest that pretrained TSFM on non-clinical domain data might not share much knowledge with clinical domain, thus prone to overfitting when full fine-tuned on small-scale clinical datasets. Further experiments regarding their generalization across datasets are available in Table 6.

4.6. Computation efficiency and model size

In the meantime, we reported the computational efficiency and parameter comparisons for different models in Figure 6. In clinical settings, it would be ideal to achieve the compromise between time cost, and quantitative results, aiming for the top-left region of Figure 6. As shown in Figure 6, whilst FMs benefit from model scaling and large-scale pretraining on non-clinical time series data (Liang et al., 2024), their performance in the clinical domain remains less impressive. Given the real-time requirements for continuous "track and trigger" monitoring, especially for critically ill patients, further research is needed to develop efficient yet powerful FM tools.

5. Conclusion

Time series forecasting is a well-studied field that has undergone significant paradigm shifts, evolving from traditional statistical methods to deep learning approaches, and more recently, to foundation models. In particular, TSFMs benefit from pretraining on large-scale time series datasets and have demonstrated exceptional performance in a variety of benchmarks, even in zero-shot learning settings. However, their performance on continuously recorded vital sign data in healthcare settings, as well as their practical applications in this domain, remain unexplored.

In this paper, we firstly investigated the forecasting potential of time series models especially TSFMs on continuously recorded vital sign data in healthcare settings. Specifically, by reorienting time series forecasting tasks toward practical clinical challenges, we integrated a pipeline that associates forecasted vital sign values with clinical deterioration tasks, complementing traditional general forecasting metrics. Through extensive experiments across different models, settings, and computational efficiency, we argue that there is substantial potential for improvement in this area. Future work should particularly focus on developing clinical TSFMs pretrained on large-scale clinical data, while also enhancing computational efficiency and incorporating multi-modal information for more accurate and efficient forecasting.

Acknowledgments

DAC was supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Wellcome Trust funded VITAL project (grant 204904/Z/16/Z); the EPSRC (grant EP/W031744/1); and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE).

References

Ali S Afshar, Yijun Li, Zixu Chen, Yuxuan Chen, Jae Hun Lee, Darius Irani, Aidan Crank, Digvijay Singh, Michael Kanter, Nauder Faraday, et al. An exploratory data quality analysis of time series physiologic signals using a large-scale intensive care unit database. *JAMIA open*, 4(3):ooab057, 2021.

Tariq Ibrahim Al-Shwaheen, Mehrdad Moghbel, Yuan Wen Hau, and Chia Yee Ooi. Use of learning approaches to predict clinical deterioration in patients based on various variables: a review of the literature. *Artificial Intelligence Review*, 55(2): 1055–1084, 2022.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2022.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Lei Clifton, David A Clifton, Marco AF Pimentel, Peter J Watkinson, and Lionel Tarassenko. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering*, 60(1):193–197, 2012.

Denise Bandeira da Silva, Diogo Schmidt, Cristiano André da Costa, Rodrigo da Rosa Righi, and Björn Eskofier. Deepsigns: A predictive model based on deep learning for the early detection of patient health deterioration. *Expert Systems with Applications*, 165:113905, 2021.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Abdur Rahim Mohammad Forkan, Ibrahim Khalil, and Mohammed Atiquzzaman. Visibid: A learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. *Computer Networks*, 113:244–257, 2017.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: Building a unified time series model. *arXiv preprint arXiv:2403.00131*, 2024.

Stephen Gerry, Jonathan Bedford, Oliver C Redfern, Hannah Rutter, Mae Chester-Jones, Marian Knight, Tony Kelly, and Peter J Watkinson. Development of a national maternity early warning score: centile based score development and delphi informed escalation pathways. *BMJ medicine*, 3 (1), 2024.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Xiao Gu, Fani Deligianni, Jinpei Han, Xiangyu Liu, Wei Chen, Guang-Zhong Yang, and Benny Lo. Beyond supervised learning for pervasive healthcare. *IEEE Reviews in Biomedical Engineering*, 2023.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3 (1):1–9, 2016.

John Kellett and Frank Sebat. Make vital signs great again–a call for action. *European Journal of Internal Medicine*, 45:13–19, 2017.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6555–6565, 2024.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):1–13, 2018.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting, 2024.

Farah E Shamout, Tingting Zhu, Pulkit Sharma, Peter J Watkinson, and David A Clifton. Deep interpretable early warning system for the detection of clinical deterioration. *IEEE Journal of Biomedical and Health Informatics*, 24(2):437–446, 2019.

Sameer Sundrani, Julie Chen, Boyang Tom Jin, Zahra Shakeri Hossein Abad, Pranav Rajpurkar, and David Kim. Predicting patient decompensation from continuous physiologic monitoring in the emergency department. *NPJ Digital Medicine*, 6 (1):60, 2023.

Eric J Topol. Medical forecasting, 2024.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *arXiv preprint arXiv:2407.13278*, 2024b.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

Appendix A. Dataset Preparation

A.1. Dataset Descriptions

The eICU (Pollard et al., 2018) and MIMIC-III WDB (Johnson et al., 2016) are two large publicly available (upon appropriate data usage agreement) datasets, which provide high-frequency, long-term, and continuous vital sign recordings collected from the bedside monitors in ICU settings. We focus on five typical vital signs, including heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), respiratory rate (RR), and oxygen saturation (SpO2). These vital signs are crucial indicators of intermediate health conditions and are widely used in monitoring patient status in critical care environments.

A.2. Preprocessing

The details of preprocessing are illustrated as below.
Resampling. Based on each recording, all data was resampled to a uniform frequency by calculating the median values every 5-minute interval, to ensure consistency across time.
Segment Extraction. Subsequently, we extracted segments of length $H+P$ (historical window H and prediction window P), from the long-term recordings using a sliding window with a step size of 2 hours.
Data Cleaning. Segments containing more than 1 consecutive hour of missing data, consistently flatten signals, as well as negative values, were filtered out to ensure high-quality data for analysis.
Missing Data Imputation. We applied forward filling for imputation of the remaining time series where small-interval data missingness occurred. If forward filling was not possible (e.g., at the start of a recording), the gaps were filled using median values calculated from the available data.

It should be noted that MIMIC-III includes multiple recordings from different ICU admissions for certain subjects. In practice, we consider these as recordings from different subjects, during data split procedures.

Appendix B. Code Implementations

B.1. Conventional Methods

For conventional time series forecasting models, we set the embedding dimension to 64 and the batch size to 128. We conducted a grid search for the learning rate and the number of encoder layers using the sets $\{0.0001, 0.0005, 0.001\}$ and $\{1, 2, 4\}$, respectively. All other parameters were kept at their default values as specified in Time-Series-Library[1].

B.2. Time Series Foundation Models

We followed the official fine-tuning implementations for FMs Chronos[2], Moirai[3], and Lag-Llama[4]. We followed the officially recommended batch size, and if necessary, we set the batch size to accomodate the single GPU memory capacity, as listed in Table A1. In addition, we conducted a grid search for the learning rate using values of $\{0.0001, 0.0005, 0.001\}$.

Table A1: Batchsize settings of full fine-tuning for existing time series foundation models.

Methods	Chronos -S	Chronos -B	Chronos -L	Moirai -S	Moirai -B	Moirai -L	Lag -Llama
Batchsize	64	32	8	64	32	8	128

We also outline the details of the TSFM settings in Table A2. It is worth noting that existing TSFMs have limited access to health-related data, and in all cases, no access to vital sign data, during pretraining. All code implementation will be released upon acceptance.

Table A2: Time series foundation model comparison.

Model	**Source**	**Multi-variate**	**Proba-bilistic**	**# Data***	**Inc. Health‡**
TimesFM	Google	✗	✗	>100B	✗
Chronos	Amazon	✗	✓	-	✗
Moirai	Salesfore	✓	✗	>27B	✓†
Lag-Llama	Morgan Stanley	✓	✓	<1B	✗

* Pretraining data size reported in Woo et al. (2024).
‡ Whether including (inc.) health data for pretraining.
† Including Hospital, COVID Deaths, US Births.

Appendix C. Explanation of discrepancy of different metrics

Here, we provide an intuitive explanation of the differences between general time series metrics and clinical deterioration related metrics, in Figure A1. With

1. https://github.com/thuml/Time-Series-Library
2. https://github.com/amazon-science/chronos-forecasting
3. https://github.com/SalesforceAIResearch/uni2ts/
4. https://github.com/time-series-foundation-models/lag-llama

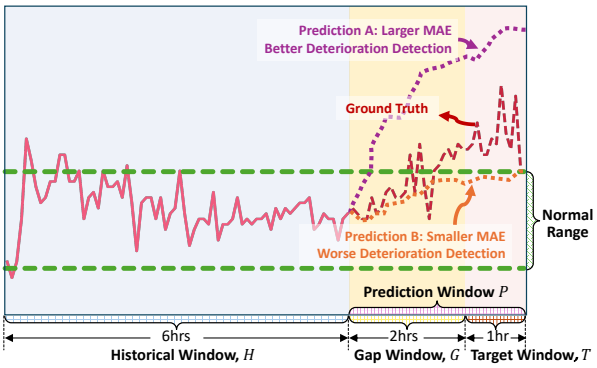


Figure A1: Illustration of the comparison of different metrics. Based on the historical window as input, we compare two different forecasted values Prediction A and Prediction B, against Ground Truth, in terms of different metrics. Although Prediction A shows a higher MAE than Prediction B, it actually demonstrates better performance in detecting deterioration.

the historical window as input, we evaluate two forecasted values, Prediction A and Prediction B, in comparison to the Ground Truth, using different metrics. Despite the fact that Prediction A having a higher MAE than Prediction B, it performs better in detecting deterioration.

Appendix D. Other Results

D.1. Other quantitative performance results

Additional results mentioned in the main paper, regarding different metrics, are provided here. These include the MSE metrics Table A3, recall of abnormal vital sign detection Table A4, precision of abnormal vital sign detection Table A5.

D.2. Other historical window length results

The complete results of the impact of hostrical window length, across vital sign types and datasets, are presented in Figures A2 and A3.

D.3. Other univariate *versus* multivariate results

More results of the effects of incorporating additional types of vital signs, on the forecasting performance, are presented in Figures A4 and A5.

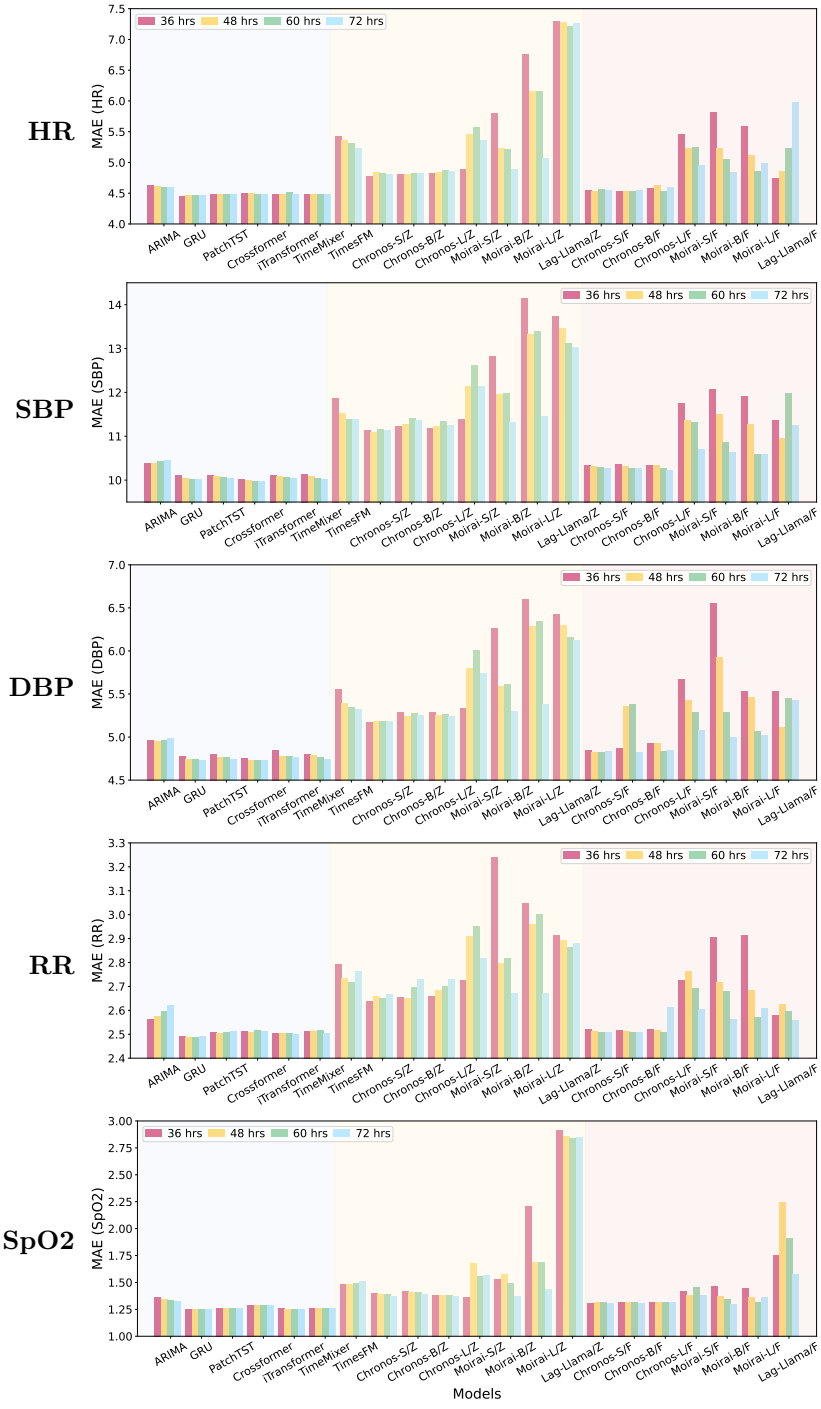


Figure A2: Performance comparison of models with different horizon lengths for vital sign forecasting on eICU dataset. "/Z" and "/F" denote zero-shot and full fine-tuning settings.

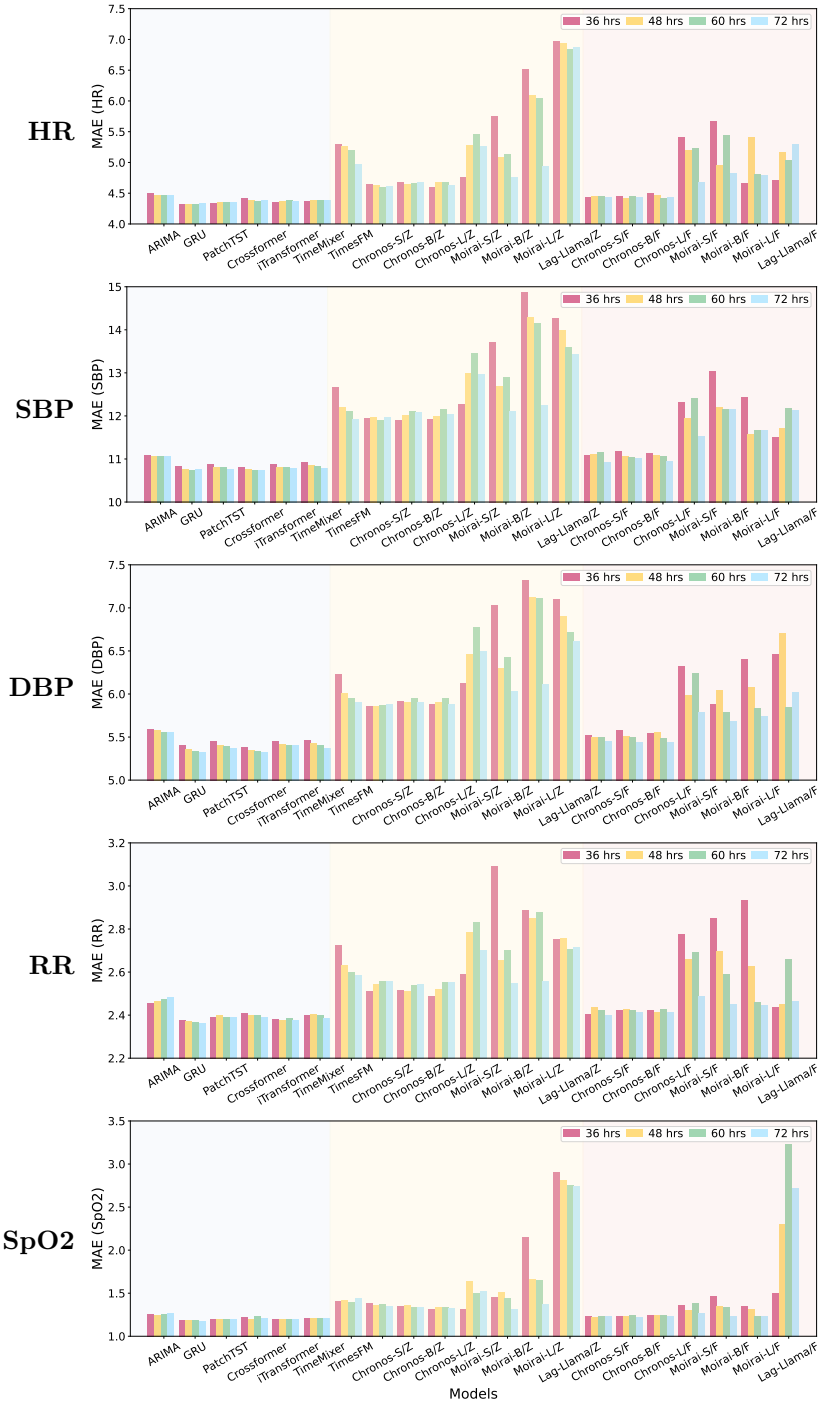
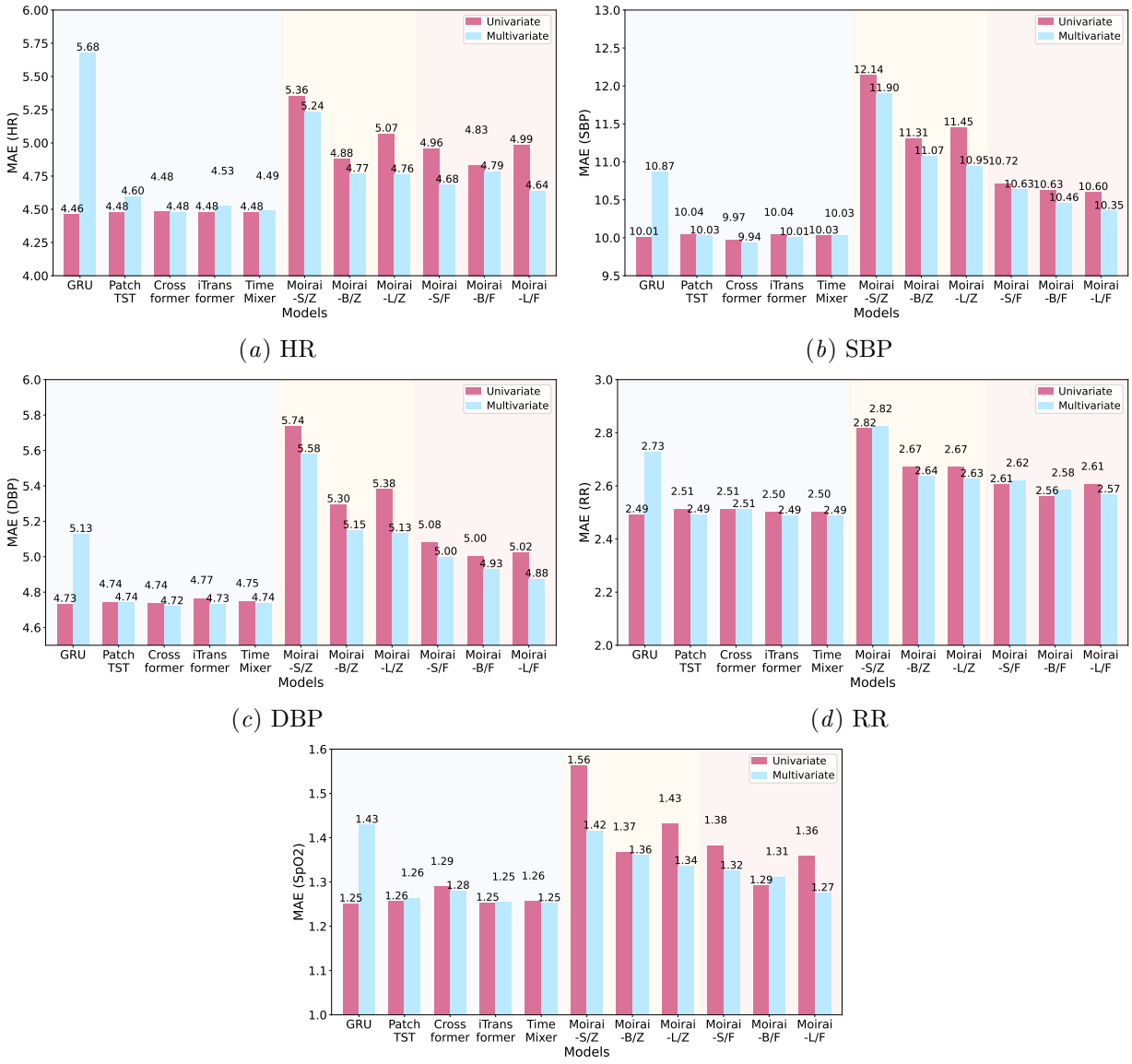


Figure A3: Performance comparison of models with different horizon lengths for vital sign forecasting on MIMIC-III dataset. "/Z" and "/F" denote zero-shot and full fine-tuning settings.



(a) HR

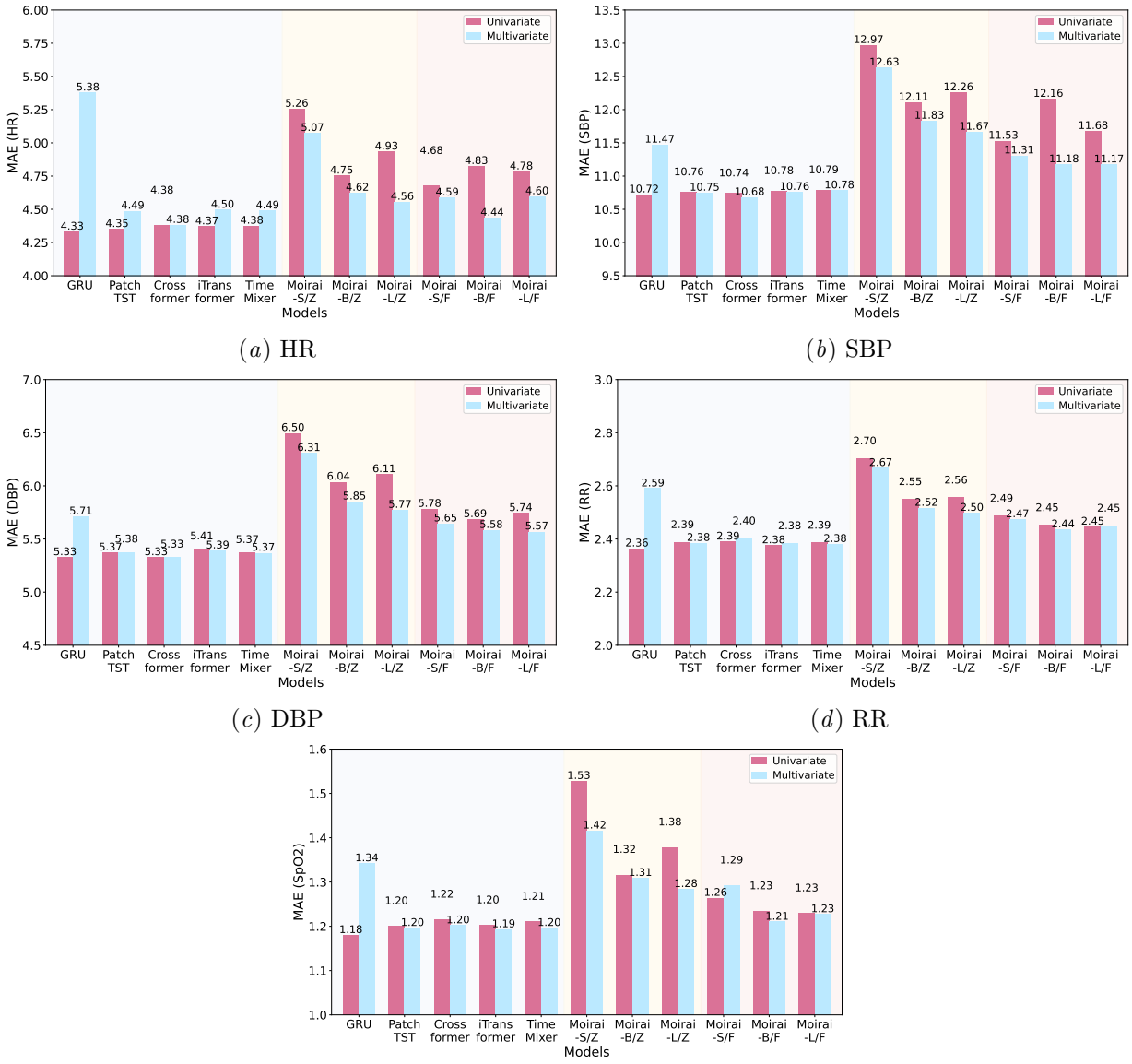
(b) SBP

(c) DBP

(d) RR

(e) SpO2

Figure A4: Performance comparison of models training in univariate and multivariate settings for vital sign forecasting on eICU dataset. "/Z" and "/F" denote zero-shot and full fine-tuning settings.



(a) HR

(b) SBP

(c) DBP

(d) RR

(e) SpO2

Figure A5: Performance comparison of models training in univariate and multivariate settings for vital sign forecasting on MIMIC-III dataset. "/Z" and "/F" denote zero-shot and full fine-tuning settings.

Table A3: Vital sign forecasting performance measured with Mean Square Error (MSE). **Bold** indicates the best result, and Underline indicates the second best result in each group.

Setting	Methods	eICU					MIMIC-III				
		HR	SBP	DBP	RR	SpO2	HR	SBP	DBP	RR	SpO2
Training from Scratch	**ARIMA**	56.167	219.236	67.571	17.555	5.511	53.561	237.740	67.155	14.338	5.498
	GRU	49.574	207.648	**64.261**	16.520	**4.592**	47.122	228.365	64.136	13.524	5.029
	PatchTST	49.701	207.544	64.543	16.585	4.608	46.684	227.452	64.357	13.472	5.113
	Crossformer	**49.087**	**203.623**	65.926	**16.228**	4.716	**46.190**	**224.444**	**63.018**	**12.868**	**4.911**
	iTransformer	50.489	209.221	64.868	16.685	4.612	47.976	229.750	65.145	13.573	5.111
	TimeMixer	49.982	208.026	64.325	16.613	4.619	47.306	229.048	64.454	13.546	5.090
Zero-Shot	**TimesFM**	66.202	266.394	**75.950**	20.249	7.008	60.739	**288.344**	**76.778**	**15.953**	8.294
	Chronos-Small	**60.166**	**263.099**	80.507	20.351	**5.352**	**54.929**	297.728	79.351	17.143	7.304
	Chronos-Base	60.806	277.250	83.189	21.096	5.489	56.974	303.741	80.170	16.882	6.267
	Chronos-Large	61.893	270.454	82.875	20.851	5.378	56.693	299.266	79.446	17.099	6.079
	Moirai-Small	70.364	303.343	88.538	20.981	6.973	68.557	337.480	91.682	18.042	7.839
	Moirai-Base	60.702	267.980	78.674	**19.713**	5.457	58.674	298.441	82.280	16.249	**5.259**
	Moirai-Large	65.459	276.508	81.483	20.118	6.110	63.027	310.308	83.955	16.732	6.905
	Lag-Llama	103.634	308.255	91.826	22.196	18.562	95.516	324.182	87.541	17.614	17.625
Full Fine-Tuning	**Chronos-Small**	53.056	220.174	68.955	18.034	5.130	49.671	238.996	68.224	15.077	5.305
	Chronos-Base	**53.010**	220.521	**68.562**	17.994	**4.999**	**49.350**	243.934	**67.769**	15.177	**5.261**
	Chronos-Large	53.530	**218.791**	69.974	18.801	5.141	49.878	**243.275**	67.878	14.862	5.326
	Moirai-Small	60.305	233.253	72.268	18.168	5.424	54.140	259.880	73.644	15.138	5.609
	Moirai-Base	57.969	231.238	71.441	**17.874**	5.008	54.854	288.718	72.426	14.778	5.429
	Moirai-Large	58.258	230.797	70.061	18.309	5.542	55.153	266.333	73.397	**14.548**	5.380
	Lag-Llama	92.463	261.706	82.553	18.043	6.179	66.490	304.727	82.920	14.629	14.404

Table A4: Physiological decompensation detection performance measured with recall. **Bold** indicates the best result, and Underline indicates the second best result in each group.

Setting	Methods	eICU					MIMIC-III				
		HR	SBP	DBP	RR	SpO2	HR	SBP	DBP	RR	SpO2
Training from Scratch	**ARIMA**	**0.682**	0.569	0.558	0.61	**0.685**	**0.662**	0.517	0.546	0.592	0.568
	GRU	0.598	0.598	0.598	0.626	0.634	0.560	0.551	0.559	0.615	**0.621**
	PatchTST	0.581	**0.627**	**0.612**	**0.659**	0.658	0.521	**0.586**	0.566	0.592	0.577
	Crossformer	0.594	0.588	0.607	0.626	0.624	0.562	0.529	0.555	0.607	0.601
	iTransformer	0.604	0.588	0.605	0.620	0.632	0.569	0.551	0.565	0.609	0.573
	TimeMixer	0.619	0.594	0.607	0.636	0.631	0.575	0.538	**0.567**	**0.617**	0.584
Zero-Shot	**TimesFM**	0.560	0.568	0.541	**0.579**	0.556	0.524	**0.590**	0.556	0.564	0.506
	Chronos-Small	0.598	**0.578**	0.601	0.571	0.570	0.585	0.552	**0.578**	**0.567**	0.601
	Chronos-Base	**0.601**	0.577	0.611	0.552	**0.586**	**0.635**	0.566	0.577	0.566	0.608
	Chronos-Large	0.587	0.568	**0.613**	0.561	0.551	0.614	0.586	0.554	0.564	0.548
	Moirai-Small	0.552	0.544	0.537	0.566	0.558	0.519	0.522	0.535	0.532	0.549
	Moirai-Base	0.549	0.542	0.570	0.569	0.584	0.541	0.539	0.548	0.536	**0.614**
	Moirai-Large	0.548	0.543	0.555	0.568	0.521	0.546	0.509	0.539	0.524	0.541
	Lag-Llama	0.499	0.506	0.514	0.516	0.503	0.500	0.498	0.510	0.516	0.498
Full Fine-Tuning	**Chronos-Small**	0.579	0.573	0.599	0.591	0.548	**0.561**	0.532	0.524	0.570	0.495
	Chronos-Base	**0.582**	0.566	0.608	**0.595**	**0.555**	0.560	0.537	0.538	**0.588**	0.496
	Chronos-Large	0.581	0.572	0.604	0.566	**0.555**	0.547	**0.556**	0.550	0.581	0.510
	Moirai-Small	0.563	0.530	0.544	0.580	0.551	0.528	0.510	0.522	0.544	0.507
	Moirai-Base	0.513	0.527	0.561	0.561	0.548	0.521	0.501	0.527	0.524	**0.511**
	Moirai-Large	0.523	0.534	0.522	0.546	0.513	0.513	0.502	0.530	0.529	**0.511**
	Lag-Llama	0.534	**0.596**	**0.647**	0.539	0.504	0.553	0.498	**0.581**	0.567	0.392

Table A5: Physiological decompensation detection performance measured with precision. **Bold** indicates the best result, and Underline indicates the second best result in each group.

Setting	Methods	eICU					MIMIC-III				
		HR	**SBP**	**DBP**	**RR**	**SpO2**	**HR**	**SBP**	**DBP**	**RR**	**SpO2**
Training from Scratch	**ARIMA**	0.614	0.602	0.590	0.545	0.558	0.598	0.535	0.595	0.588	0.565
	GRU	0.662	**0.606**	0.625	0.551	**0.580**	0.628	0.577	0.591	**0.614**	**0.614**
	PatchTST	**0.663**	0.595	**0.636**	0.552	0.566	0.674	**0.608**	0.604	0.562	0.579
	Crossformer	0.659	0.600	0.623	**0.554**	0.551	0.668	0.555	0.594	0.611	**0.614**
	iTransformer	0.658	0.599	0.626	0.551	0.578	**0.680**	0.579	0.599	0.605	0.607
	TimeMixer	0.649	0.593	0.623	0.552	0.549	0.641	0.559	**0.608**	0.597	0.558
Zero-Shot	**TimesFM**	0.617	0.572	0.547	0.534	0.520	0.564	**0.604**	0.579	0.558	0.502
	Chronos-Small	0.633	0.570	0.608	0.544	0.520	0.597	0.552	0.565	0.559	0.532
	Chronos-Base	0.629	0.560	0.591	0.538	0.524	0.635	0.558	0.570	0.564	0.534
	Chronos-Large	0.631	0.559	0.601	0.542	0.515	0.626	0.575	0.547	0.566	0.515
	Moirai-Small	0.638	0.596	0.589	0.563	0.537	0.587	0.562	**0.626**	0.561	0.539
	Moirai-Base	0.631	**0.600**	**0.613**	**0.572**	**0.539**	**0.664**	0.592	0.582	**0.585**	**0.559**
	Moirai-Large	**0.645**	0.599	0.607	0.557	0.507	0.600	0.526	0.581	0.556	0.515
	Lag-Llama	0.483	0.596	0.571	0.518	0.527	0.486	0.468	0.545	0.539	0.486
Full Fine-Tuning	**Chronos-Small**	0.638	0.608	0.616	0.560	0.517	0.649	0.573	0.545	0.560	0.497
	Chronos-Base	0.648	0.589	0.615	0.560	0.519	0.624	0.579	0.575	0.585	0.498
	Chronos-Large	0.644	0.613	0.620	0.549	0.519	0.617	**0.609**	**0.591**	0.569	0.505
	Moirai-Small	0.629	0.610	0.611	**0.574**	0.536	0.669	0.534	0.583	**0.626**	0.506
	Moirai-Base	0.646	0.625	**0.643**	0.565	**0.578**	**0.737**	0.503	0.583	0.579	**0.539**
	Moirai-Large	**0.674**	**0.640**	0.639	0.566	0.511	0.598	0.505	0.579	0.565	0.534
	Lag-Llama	0.531	0.568	0.562	0.553	0.507	0.554	0.497	0.550	0.549	0.482