Towards Large-scale Clinical Multi-variate Time-series Datasets

Manuel Burger[*†]
manuel.burger@inf.ethz.ch

Fedor Sergeev[*†]
fedor.sergeev@inf.ethz.ch

Malte Londschien[‡§†] **Daphné Chopard**[†¶] **Hugo Yèche**[†] **Eike Gerdes**[‖]

Polina Leshetkina[**] **Alexander Morgenroth**[†] **Zeynep Babür**[††] **Jasmina Bogojeska**[††]

Martin Faltys[‡‡]

Rita Kuznetsova[*†]
rita.kuznetsova@inf.ethz.ch

Gunnar Rätsch[*†]
raetsch@inf.ethz.ch

Abstract

Notable progress has been made in generalist medical large language models across various healthcare areas. However, large-scale modeling of in-hospital time series data - such as vital signs, lab results, and treatments in critical care - remains underexplored. Existing datasets are relatively small, but combining them can enhance patient diversity and improve model robustness. To effectively utilize these combined datasets for large-scale modeling, it is essential to address the distribution shifts caused by varying treatment policies, necessitating the harmonization of treatment variables across the different datasets. This work aims to establish a foundation for training large-scale multi-variate time series models on critical care data and to provide a benchmark for machine learning models in transfer learning across hospitals to study and address distribution shift challenges. We introduce a harmonized dataset for sequence modeling and transfer learning research, representing the first large-scale collection to include core treatment variables. Future plans involve expanding this dataset to support further advancements in transfer learning and the development of scalable, generalizable models for critical healthcare applications.

[*] equal contribution

[†] Department of Computer Science, ETH Zurich, Switzerland

[‡] Seminar for Statistics, ETH Zurich, Switzerland

[§] AI Center, ETH Zurich, Switzerland

[¶] Department of Intensive Care and Neonatology and Children's Research Center, University Children's Hospital Zurich, University of Zurich, Switzerland

[‖] University of Zurich, Switzerland

[**] Department of Health Science and Medicine, University of Luzern, Switzerland

[††] Zurich University of Applied Sciences (ZHAW), Switzerland

[‡‡] Department of Intensive Care Medicine, University Hospital and University of Bern, Switzerland

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

1 Introduction

Foundation models trained on complex multi-modal medical data have the potential to significantly transform healthcare [33]. Considerable advancements have been made in the development of generalist medical Large Language Models (LLM) [43, 6], computer vision models in pathology [51], single-cell multi-omics models [8], and sequence models on coded Electronic Health Records (EHR) [54].

One area that remains underexplored is the foundation models for critical care time series[10]. It is promising because of the prospective benefits for patients and the availability of large (multi-site and multi-national) and rich (multi-variate, including vital signs, lab measurements, and treatments) data.

Developing a large-scale foundation model with robust generalization capabilities across hospitals and countries requires a comprehensive dataset with high patient diversity [40]. Individually published datasets from Intensive Care Units (ICU) and Emergency Departments (ED) [26, 12, 46, 41, 25, 53] are relatively small compared to modern standards in fields such as Natural Language Processing (NLP) [13]. However, by aggregating them, it is possible to scale the number of admissions by an order of magnitude and increase their diversity. Previous works addressing such aggregation did not include all the ICU datasets, add ED datasets, or harmonize treatment variables [4, 57].

A key challenge in creating a foundation model is to ensure its robustness to distribution shifts. In the clinical domain, this is especially difficult because of substantial differences in recording formats and treatment policies between hospitals and countries [21]. Robustness to these shifts would suggest that the model generalizes beyond cohort-specific pattern matching and achieves a deeper understanding of human physiology. Specifically on critical care time series, most previous works considered single-center performance [19, 58, 5]. The few publications that did consider transfer, either focused on a specific task [34] or did not attempt to improve model generalization and ensure its robustness [48].

Our aim is to establish the foundation for training and evaluating large-scale multi-variate time-series models on real-world hospital data from critical care. To achieve this goal, we create a large multi-center dataset covering a wide array of clinical features and build an understanding of what machine learning algorithms work well on such data.

We expect this work to become the basis for a future foundational model with a wide range of downstream medical applications. Specifically, it will unlock research for small cohorts of specific patients using few-shot learning or fine-tuning, mirroring the impact of pretrained language models in NLP. Furthermore, for the ML community, the dataset we present will be a valuable resource for research into sequence modeling, meta and transfer learning, domain adaptation, and generalization.

Our current contributions are two-fold:

- *Dataset*. We introduce the largest harmonized critical care time series medical dataset. It is the first of such datasets to (a) harmonize the core treatment variables, (b) include datasets from both ICU and ED, (c) incorporate data from Asia in addition to Europe and the USA, and (d) provide annotations and results on multiple organ failure tasks on the same data. It is extendable and can be used for research in sequence modeling, domain generalization, and meta-learning.
- *Benchmark*. We run a comprehensive benchmark of machine learning models on the new dataset. We perform transfer studies and evaluate performance on clinically relevant real-time prediction tasks in-distribution as well as out-of-distribution.

2 Experiments

2.1 Setup

Data With the goal of maximizing the number of harmonized physiological measurements and treatment data points, we incorporate all ICU datasets that are freely available to the academic community. These include datasets from the USA (MIMIC-III [24], MIMIC-IV [27, 24], and

[10] We call *critical care* a setting, where a patient is being closely monitored (e.g., in emergency departments and intensive care units, during surgery, etc.)

eICU [38]), Europe (AmsterdamUMCdb [46], SICdb [41], and HiRID [12]), and China (PICdb [30] and Zigong EHR [56]). Additionally, we incorporate an ED dataset [25]. Most datasets are available on the Physionet Platform [14] or directly with the dataset provider (e.g. AmsterdamUMCdb [46]). Further details are in Appendix C.

Figure 1 shows visualization of harmonized and processed data by t-SNE [49]. The apparent clustering by source hospital emphasizes the challenge of developing a predictor that is robust to these distribution shifts across sites.

Models We consider two groups of machine learning algorithms. The first group consists of classical machine learning methods (LightGBM [28] for gradient boosted decision trees and regularized Linear Regression [39]), which are highly effective for real-time prediction tasks on critical care time series [19, 22, 58]. For these models, we either use the forward-filled last available measurement for each variable (*Last Meas.)* or include hand-extracted features from the history based on the work by Soenksen et al. [44], which we further expanded to improve performance (Appendix C.4.3). The second group is focused on deep learning methods. We select established and state-of-the-art sequence architectures for this group: Gated Recurrent Unit (GRU) [7], Transformer [50], Mamba [9] and xLSTM [3]. Training details in Appendix D.

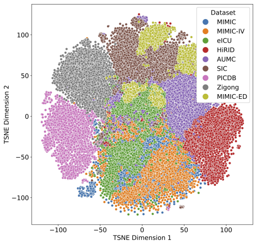


Figure 1: Visualization of harmonized and processed data by t-SNE [49]. Each point represents a time step.

2.2 In-distribution and out-of-distribution Benchmark

Single-center in-distribution training represents the classical setting where a model is trained and evaluated on the train and test subsets of a single source dataset. We can furthermore, consider training on multiple datasets jointly and reporting individual in-distribution performances. Comprehensive benchmark results can be found in Tables 1 and 2 in Appendix A. Results for disposition prediction on MIMIC-IV-ED are presented in Table 5 (Appendix A).

In the single-center out-of-distribution setting, we train a model on any single dataset and report test performance on the held-out dataset (e.g. Figure 2 shows results for LightGBM, further results on other models in Appendix A.3). In the multi-center hold-out setting, we train a model on all but the target dataset and then report test set performance on the held-out dataset.

Results for out-of-distribution transfer for single- and multi-center training are presented in Tables 3 and 4 (Appendix A). Hyperparameter optimization is always performed on the in-distribution validation sets corresponding to the collection of training sets.

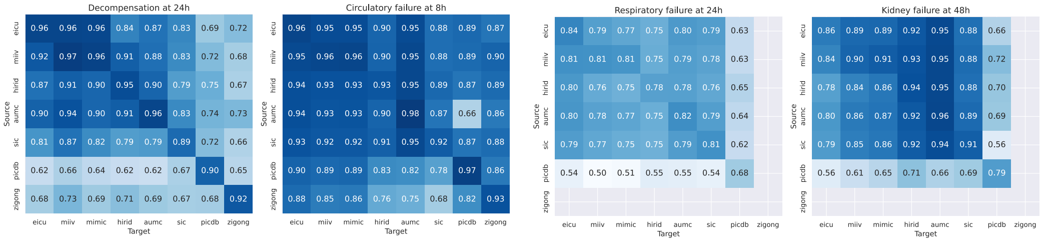


Figure 2: LGBM (Feat.) single-center transfer performance heatmaps (AUROC).

2.3 Fine-tuning Study

We performed a supervised pretraining and fine-tuning study, as shown in Figure 3 and Appendix A Figure 5, using the HiRID dataset [12]. We trained models from scratch on progressively increasing number of HiRID patients: LightGBM with extracted features, GRU as the best performing deep sequence architecture, and Mamba as a modern RNN variant. Further, we pretrained a GRU (or Mamba) backbone on all other datasets in a supervised fashion and reported the zero-shot transfer performance without using any HiRID data. Finally, we initialized a GRU (or Mamba) network with the aformentioned supervised pretrained weights and fine-tuned either the full network or only the linear logit head.

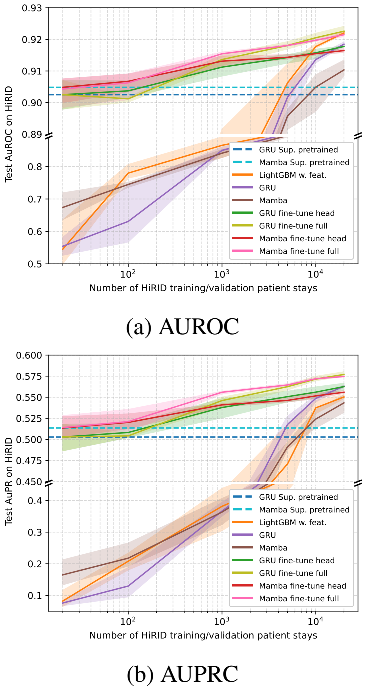
3 Discussion

Overall, the transfer performance in- and out-of-distribution suggests that even without fine-tuning the resulting time series models are capable of performing early event prediction for relevant medical labels reasonably well. In all settings, we see that gradient-boosted trees with feature extraction are the best-performing model across the board (see Tables 1 and 3). This is consistent with the previous findings [22, 58].

At the same time, we note that the performance of deep models is often within just one or two AUROC points of the classical algorithms. For disposition prediction, the gap is even smaller, around a tenth of a point (Table 5). In some cases, they manage to outperform tree-based alternatives, especially in multi-center settings.

From the transfer heatmaps Figure 2 we notice that (1) models generally transfer better in particular groups of datasets and (2) the transfer performance depends on the task.

On average, both LGBM and GRU generally transfer well between the eICU, MIMIC, HiRID, UMCdb, and SICdb datasets (Figures 4a and 4b). We hypothesize that this is due to their locality. These datasets originate from the USA and Europe, where clinical practices might be more similar than for example between the USA and China (PICdb and Zigong EHR datasets). This is reinforced by the particularly good transfer between eICU and MIMIC, both originating in the USA.



(a) AUROC

(b) AUPRC

Figure 3: Supervised fine-tuning study performed on HiRID for circulatory failure predictions at 8 hour horizon by progressively increasing the number of patients used for training or fine-tuning.

The fine-tuning study (see Figure 3 and Appendix A Figure 5) suggests that the models trained on the harmonized collection of datasets are able to generalize to a new, previously unseen dataset. They outperform a model trained from scratch for dataset sizes of up to tens of thousands of admissions. The practical implication is that training on publicly available datasets should be the go-to strategy for small and medium scale studies on critical care time series. We also see that a fine-tuned GRU model performs better or is on par with the LGBM model trained from scratch on admission counts fewer and larger than 10,000. This result suggests that deep learning models might be a preferable choice when transferred from large datasets.

4 Conclusion

In this work, we established the foundation for large-scale time series models on critical care data. We created the largest harmonized dataset that includes hospitals from three continents, incorporates treatment variables, and integrates data from both ICU and ED units. The dataset is further supported by a comprehensive transfer learning benchmark.

Our results demonstrated that with the access to an increased amount of carefully harmonized and labeled data, machine learning models are capable of generalizing when transferring across countries and continents, even without extensive fine-tuning. Notably, gradient-boosted trees with feature extraction consistently outperformed other models, although deep learning models came remarkably close, particularly in multi-center settings and for disposition prediction tasks. Importantly, our study highlights how dataset resolution and geographic origin influence transferability. Finally, fine-tuned models, trained on harmonized datasets, significantly improved performance on previously unseen data, especially for small and medium-sized datasets.

In future work, we plan to explore further improvements to the data coverage and training procedure to create the first foundation model for critical care time series and beyond.

Acknowledgments

This work was supported as part of the Swiss AI Initiative (swiss-ai.org) by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a02 on Alps. Computational data analysis was performed at Leonhard Med (https://sis.id.ethz.ch/services/sensitiveresearchdata/) secure trusted research environment at ETH Zurich.

The project was supported by grant #2022-278 of the Strategic Focus Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain (Swiss Federal Institutes of Technology) and ETH Core funding (to G.R.).

Malte Londschien was supported by the ETH Foundations of Data Science and the ETH AI Center. Fedor Sergeev was supported by grant #902 of the Strategic Focus Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain (Swiss Federal Institutes of Technology). Fedor Sergeev thanks Vincent Fortuin for his helpful suggestions for improving the manuscript. Daphné Chopard received funding from grant #2021-911 of the Strategic Focal Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain (Swiss Federal Institutes of Technology).

References

[1] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[2] B. Arnrich, E. Choi, J. A. Fries, M. B. McDermott, J. Oh, T. J. Pollard, N. Shah, E. Steinberg, M. Wornow, and R. van de Water. Medical event data standard (meds): Facilitating machine learning for health. 2024. Accepted as a Workshop Paper at TS4H@ICLR2024.

[3] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

[4] N. Bennett, D. Plečko, I.-F. Ukor, N. Meinshausen, and P. Bühlmann. ricu: R’s interface to intensive care data. *GigaScience*, 12:giad041, 2023.

[5] E. Chen, A. Kansal, J. Chen, B. T. Jin, J. Reisler, D. E. Kim, and P. Rajpurkar. Multimodal clinical benchmark for emergency care (mc-bec): A comprehensive benchmark for evaluating foundation models in emergency medicine. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Z. Chen, A. Hernández-Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

[7] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8): 1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

[9] T. Dao and A. Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL https://arxiv.org/abs/2405.21060.

[10] A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[11] W. Falcon and The PyTorch Lightning team. PyTorch Lightning, Mar. 2019. URL https://github.com/Lightning-AI/lightning.

[12] M. Faltys, M. Zimmermann, X. Lyu, M. Hüser, S. Hyland, G. Rätsch, and T. Merz. HiRID, a high time-resolution icu dataset (version 1.1.1). *PhysioNet*, 2021. doi: 10.13026/nkwc-js72. URL `https://doi.org/10.13026/nkwc-js72`.

[13] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL `https://arxiv.org/abs/2101.00027`.

[14] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220, 2000.

[15] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data, 2023. URL `https://arxiv.org/abs/2106.11959`.

[16] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[17] L. L. Guo, E. Steinberg, S. L. Fleming, J. Posada, J. Lemmon, S. R. Pfohl, N. Shah, J. Fries, and L. Sung. Ehr foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(1):3767, 2023.

[18] L. L. Guo, J. Fries, E. Steinberg, S. L. Fleming, K. Morse, C. Aftandilian, J. Posada, N. Shah, and L. Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7(1):171, 2024.

[19] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[20] M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt. Set functions for time series, 2020. URL `https://arxiv.org/abs/1909.12064`.

[21] M. Hüser, X. Lyu, M. Faltys, A. Pace, M. Hoche, S. Hyland, H. Yèche, M. Burger, T. M. Merz, and G. Rätsch. A comprehensive ml-based respiratory monitoring system for physiological monitoring & resource planning in the icu. *medRxiv*, 2024. doi: 10.1101/2024.01.23.24301516. URL `https://www.medrxiv.org/content/early/2024/01/23/2024.01.23.24301516`.

[22] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

[23] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV" (version 2.2). PhysioNet (2023). URL `https://physionet.org/content/mimiciv/2.2/`.

[24] A. Johnson, T. Pollard, and M. Roger. MIMIC-III Clinical Database, 2016. URL `https://physionet.org/content/mimiciii/1.4/`.

[25] A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, S. Horng, and R. Mark. Mimic-iv-ed demo, 2023. URL `https://doi.org/10.13026/jzz5-vs76`.

[26] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[27] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[28] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[29] S.-Y. Lee, R. B. Chinnam, E. Dalkiran, S. Krupp, and M. Nauss. Prediction of emergency department patient disposition decision for proactive resource allocation for admission. *Health care management science*, 23:339–359, 2020.

[30] H. Li, X. Zeng, and G. Yu. Paediatric intensive care database, 2019.

[31] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

[32] X. Lyu, B. Fan, M. Hüser, P. Hartout, T. Gumbsch, M. Faltys, T. M. Merz, G. Rätsch, and K. Borgwardt. An empirical study on kdigo-defined acute kidney injury prediction in the intensive care unit. *Bioinformatics*, 40(Supplement_1):i247–i256, 2024.

[33] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[34] M. Moor, N. Bennett, D. Plečko, M. Horn, B. Rieck, N. Meinshausen, P. Bühlmann, and K. Borgwardt. Predicting sepsis using deep learning across international sites: a retrospective development and validation study. *EClinicalMedicine*, 62, 2023.

[35] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics - Measuring Reproducibility in PyTorch, Feb. 2022. URL `https://github.com/Lightning-AI/torchmetrics`.

[36] M. Oliver, J. Allyn, R. Carencotte, N. Allou, and C. Ferdynus. Introducing the blendedicu dataset, the first harmonized, international intensive care dataset. *Journal of Biomedical Informatics*, 146:104502, 2023. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2023.104502. URL `https://www.sciencedirect.com/science/article/pii/S153204642300223X`.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[38] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[39] I. QuantCo. glum. `https://github.com/Quantco/glum`, 2020.

[40] P. Rockenschaub, A. Hilbert, T. Kossen, P. Elbers, F. von Dincklage, V. I. Madai, and D. Frey. The impact of multi-institution datasets on the generalizability of machine learning prediction models in the icu. *Critical Care Medicine*, pages 10–1097, 2024.

[41] N. Rodemund, A. Kokoefer, B. Wernly, and C. Cozowicz. Salzburg intensive care database (sicdb), a freely accessible intensive care database. *PhysioNet https://doi. org/10.13026/ezs8-6v88*, 2023.

[42] S. Sheikhalishahi, V. Balaraman, and V. Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.

[43] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023.

[44] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(1):149, Sep 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00689-4. URL `https://doi.org/10.1038/s41746-022-00689-4`.

[45] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, 2020.

[46] P. J. Thoral, J. M. Peppink, R. H. Driessen, E. J. Sijbrands, E. J. Kompanje, L. Kaplan, H. Bailey, J. Kesecioglu, M. Cecconi, M. Churpek, et al. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the amsterdam university medical centers database (amsterdamumcdb) example. *Critical care medicine*, 49(6):e563–e577, 2021.

[47] S. Tipirneni and C. K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series, 2022. URL `https://arxiv.org/abs/2107.14293`.

[48] R. Van De Water, H. Schmidt, P. Elbers, P. Thoral, B. Arnrich, and P. Rockenschaub. Yet another icu benchmark: A flexible multi-center framework for clinical ml. *arXiv preprint arXiv:2306.05109*, 2023.

[49] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[51] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, S. Liu, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, N. Fusi, P. Mathieu, A. van Eck, D. Lee, J. Viret, E. Robert, Y. K. Wang, J. D. Kunz, M. C. H. Lee, J. Bernhard, R. A. Godrich, G. Oakley, E. Millar, M. Hanna, J. Retamero, W. A. Moye, R. Yousfi, C. Kanan, D. Klimstra, B. Rothrock, and T. J. Fuchs. Virchow: A million-slide digital pathology foundation model, 2024. URL `https://arxiv.org/abs/2309.07778`.

[52] S. Wang, M. B. A. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann. Mimic-extract: a data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, ACM CHIL '20. ACM, Apr. 2020. doi: 10.1145/3368555.3384469. URL `http://dx.doi.org/10.1145/3368555.3384469`.

[53] M. Wornow, R. Thapa, E. Steinberg, J. Fries, and N. Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

[54] M. Wornow, R. Thapa, E. Steinberg, J. A. Fries, and N. H. Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models, 2023. URL `https://arxiv.org/abs/2307.02028`.

[55] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.

[56] P. Xu, L. Chen, and Z. Zhang. Critical care database comprising patients with infection at zigong fourth people's hospital (version 1.1), 2022.

[57] C. Yang, Z. Wu, P. Jiang, Z. Lin, J. Gao, B. P. Danek, and J. Sun. Pyhealth: A deep learning toolkit for healthcare applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5788–5789, 2023.

[58] H. Yèche, R. Kuznetsova, M. Zimmermann, M. Hüser, X. Lyu, M. Faltys, and G. Rätsch. Hirid-icu-benchmark — a comprehensive machine learning benchmark on high-resolution icu data. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/5878a7ab84fb43402106c575658472fa-Paper-round1.pdf`.

[59] H. Yèche, M. Burger, D. Veshchezerova, and G. Rätsch. Dynamic survival analysis for early event prediction, 2024. URL `https://arxiv.org/abs/2403.12818`.

[60] T. Zhou, P. Niu, L. Sun, R. Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.

Dataset	Task	Single-Center							Multi-Center						
		LR (Last Meas.)	LGBM (Last Meas.)	LGBM (Feat.)	GRU	Transformer	Mamba	xLSTM	LR (Last Meas.)	LGBM (Last Meas.)	LGBM (Feat.)	GRU	Transformer	Mamba	xLSTM
MIMIC-IV	Dec. 24h	93.7	95.4	**97.3**	95.8	95.7	95.8	95.8	92.0	94.7	96.1	95.8	95.0	95.0	95.7
	Circ. 8h	93.5	95.2	**95.6**	94.9	94.9	94.8	95.0	92.5	94.6	95.1	94.5	94.4	94.5	94.7
	Resp. 24h	76.4	79.7	**81.1**	79.9	79.7	79.7	79.8	74.1	78.7	79.9	79.3	78.9	79.5	79.5
	Kidn. 48h	83.2	87.5	**89.8**	88.3	87.7	87.9	88.3	81.6	86.9	89.2	87.8	86.2	88.1	88.3
eICU	Dec. 24h	91.1	93.0	**95.8**	93.4	93.3	93.4	93.5	89.1	92.4	93.9	92.8	92.4	92.9	93.1
	Circ. 8h	94.4	95.6	**96.0**	95.4	95.4	95.2	95.3	93.2	95.2	95.6	94.7	94.8	94.8	94.9
	Resp. 24h	79.2	82.1	**83.6**	82.8	82.3	82.2	82.1	78.1	81.7	82.5	81.8	81.3	81.8	82.1
	Kidn. 48h	74.5	82.0	**85.6**	83.7	82.9	83.3	83.8	73.0	81.0	84.2	82.1	81.5	82.7	83.1
HiRID	Dec. 24h	93.0	93.8	94.5	**94.6**	94.0	94.4	94.3	92.4	94.4	95.1	94.3	94.1	94.4	94.5
	Circ. 8h	90.8	91.9	**92.6**	92.3	92.0	92.0	92.2	90.7	92.1	92.8	**92.6**	92.4	92.3	**92.6**
	Resp. 24h	75.2	76.6	78.1	77.2	76.9	76.6	76.8	75.1	77.1	**78.4**	78.0	77.5	77.4	77.3
	Kidn. 48h	91.2	93.0	93.7	92.3	91.1	91.9	92.1	90.7	93.4	**94.3**	93.6	93.2	93.0	93.4
UMCdb	Dec. 24h	88.9	92.3	**95.9**	92.9	91.8	92.1	92.3	88.4	92.3	95.2	93.2	92.9	93.4	93.4
	Circ. 8h	96.2	97.1	97.5	97.6	97.3	97.3	97.5	95.7	97.0	97.7	**97.8**	97.7	97.7	**97.8**
	Resp. 24h	78.8	80.4	**82.0**	81.2	80.7	80.6	80.5	78.2	80.5	82.1	81.4	81.2	80.8	80.7
	Kidn. 48h	93.4	95.1	**95.6**	94.5	94.1	93.9	94.3	93.4	95.8	96.1	95.2	95.0	95.2	95.0
SICdb	Dec. 24h	83.7	88.1	88.8	87.9	86.7	87.2	87.8	81.8	88.8	90.9	89.2	89.2	89.0	**89.6**
	Circ. 8h	88.7	90.3	91.6	91.0	90.7	90.6	90.5	**95.7**	90.3	91.7	91.3	91.1	91.1	91.3
	Resp. 24h	77.9	80.8	**81.4**	80.7	80.2	80.3	80.1	78.2	80.9	81.7	81.1	80.7	81.2	81.0
	Kidn. 48h	87.3	89.4	90.8	88.9	87.7	88.3	88.1	**93.4**	90.1	91.9	89.2	89.2	89.1	89.3
PICdb	Dec. 24h	85.3	85.6	**90.3**	87.8	88.0	87.2	87.8	70.6	87.0	88.8	83.2	81.8	84.5	84.0
	Circ. 8h	94.2	94.7	**96.8**	96.0	96.0	95.8	96.4	88.6	92.4	92.1	92.0	92.2	93.2	93.6
	Resp. 24h	**71.3**	66.3	68.5	68.4	70.7	70.2	65.9	65.9	59.5	59.3	66.1	66.2	66.8	67.3
	Kidn. 48h	73.5	**81.9**	78.9	63.5	63.8	66.5	69.0	57.4	71.8	81.2	67.5	67.1	67.0	64.5
Zigong	Dec. 24h	69.3	78.3	**92.2**	86.4	85.1	76.6	68.7	66.8	70.1	80.3	73.8	71.6	71.4	74.4
	Circ. 8h	88.2	92.3	**93.5**	88.8	86.6	84.8	84.7	89.0	90.0	89.1	86.0	84.9	86.0	86.2

Table 1: Benchmarking results in-distribution (AUROC). Bold is best in each row. Multi-center trains on all datasets together and provides the individual test performances. All results show the mean over three different random initialization, except for LR models that are trained using convex optimization.

A Results

A.1 In-distribution Transfer Table

In multi-center in-distribution setting a model is trained on all the harmonized datasets jointly and evaluated on a test set of a single dataset. By carefully normalizing the features not to depend on resolution and passing appropriate positional encoding at each time step, we can train the model in a multi-resolution fashion. Test results then report the individual performances of this single model on each source dataset separately.

A.2 Out-of-distribution Transfer Table

For single-center experiments, we report the performance of the model and dataset which transferred the best. We do not consider this to be a realistic deployment scenario, but rather a reference point for the multi-center results. It verifies whether training on a single dataset that is similar to the target dataset is better than training on a collection of harmonized datasets. In real-world, selecting such a training dataset would require either a strong evidence of "similarity" (hard to justify as we observe transfer difficulties even within the same country) or considerable validation effort (technically infeasible for the majority of hospitals).

The multi-center hold-out is a more realistic setting as it does not require any prior knowledge or evaluation sets for the selection of the training set for transfer. It represents a deployment scenario, where a hospital with little data suitable for ML training uses and adapts the model built on open-access data.

Dataset	Task	Single-Center							Multi-Center						
		LR (Last Meas.)	LGBM (Last Meas.)	LGBM (Feat.)	GRU	Transformer	Mamba	xLSTM	LR (Last Meas.)	LGBM (Last Meas.)	LGBM (Feat.)	GRU	Transformer	Mamba	xLSTM
MIMIC-IV	Dec. 24h	44.7	53.5	**62.4**	57.4	55.6	56.4	56.4	39.2	49.8	56.4	52.9	49.5	53.1	53.5
	Circ. 8h	59.7	66.8	**68.2**	65.7	65.4	65.3	66.2	55.7	63.8	66.0	63.7	63.2	63.9	64.8
	Resp. 24h	75.4	79.6	**81.2**	79.6	79.2	79.3	79.4	73.6	78.5	80.0	79.4	78.7	79.4	79.4
	Kidn. 48h	40.8	48.6	**51.8**	46.9	45.6	46.0	47.0	39.5	47.2	50.8	47.6	45.9	48.3	48.6
eICU	Dec. 24h	33.1	37.9	**51.4**	41.2	40.9	40.8	40.9	30.0	36.8	41.8	39.5	38.0	38.9	39.5
	Circ. 8h	56.8	64.7	**65.3**	63.8	63.2	62.8	63.5	54.2	61.7	63.4	60.9	60.6	61.4	62.2
	Resp. 24h	72.6	78.2	**80.0**	79.1	78.3	78.4	78.3	71.3	77.7	78.7	78.0	77.1	77.9	78.2
	Kidn. 48h	33.6	42.5	**47.3**	43.3	42.5	43.1	43.8	32.7	40.7	44.5	41.2	40.7	42.4	42.9
HiRID	Dec. 24h	43.3	50.1	52.9	54.1	51.9	52.7	53.4	42.1	51.3	**55.7**	53.4	51.8	52.5	52.7
	Circ. 8h	52.5	55.4	57.3	57.6	56.7	56.9	57.6	51.8	57.0	59.0	**59.2**	58.3	57.8	58.6
	Resp. 24h	88.7	90.1	90.9	90.3	90.1	89.9	90.0	88.3	90.4	**91.0**	90.9	90.6	90.6	90.6
	Kidn. 48h	44.4	52.6	54.6	50.1	48.8	47.9	48.7	41.4	53.8	**57.9**	50.7	52.8	49.6	51.4
UMCdb	Dec. 24h	36.0	42.2	**55.7**	47.1	44.0	43.8	43.6	35.4	43.7	54.4	50.5	48.5	48.5	48.1
	Circ. 8h	85.5	89.8	**91.7**	90.8	90.3	90.5	90.8	82.2	87.6	91.4	91.4	91.4	91.4	91.7
	Resp. 24h	85.3	86.9	**88.1**	87.5	87.0	86.9	86.8	84.6	87.0	88.1	87.7	87.5	87.3	87.1
	Kidn. 48h	50.8	56.5	58.9	56.0	53.9	53.6	54.2	49.1	59.1	**63.3**	56.6	58.5	55.2	57.2
SICdb	Dec. 24h	31.0	31.7	32.6	33.4	34.5	32.3	32.9	28.4	35.6	39.7	38.5	38.5	38.1	**40.4**
	Circ. 8h	49.1	53.4	56.2	55.1	54.0	54.3	54.1	47.8	53.9	56.6	56.6	55.8	55.8	**57.0**
	Resp. 24h	85.6	88.1	88.6	88.0	87.6	87.4	87.3	85.5	88.3	**89.0**	88.6	88.2	88.5	88.4
	Kidn. 48h	31.5	33.9	42.6	35.9	31.4	32.4	33.1	29.1	39.5	**45.8**	37.0	37.9	35.8	36.2
PICdb	Dec. 24h	15.2	12.0	16.5	16.6	16.5	16.2	16.4	6.9	16.0	**19.1**	13.7	12.0	13.5	13.2
	Circ. 8h	92.9	93.9	**96.3**	95.1	95.2	95.0	95.7	87.0	91.1	90.9	90.8	91.1	92.2	92.8
	Resp. 24h	8.2	8.2	**11.3**	10.6	10.4	11.1	9.9	4.6	3.6	3.7	4.8	5.5	5.5	7.2
	Kidn. 48h	5.1	9.2	**11.8**	3.2	3.7	3.8	3.9	6.6	7.6	11.3	9.3	8.4	7.3	7.1
Zigong	Dec. 24h	22.3	30.3	**54.0**	41.4	37.8	29.9	32.6	20.3	20.4	28.3	23.2	24.5	22.5	25.5
	Circ. 8h	98.0	98.6	**98.8**	98.0	97.5	97.0	94.4	98.2	98.3	98.2	97.6	97.2	97.3	97.5

Table 2: Benchmarking results in-distribution (AUPRC). Bold is best in each row. Multi-center trains on all datasets together and provides the individual test performances. All results show the mean over three different random initialization, except for LR models that are trained using convex optimization.

Dataset	Task	Single-Center			Multi-Center hold-out						
		AUROC	Best Model	Src. Data	LR (Last Meas.)	LGBM (Last Meas.)	LGBM (Feat.)	GRU	Transformer	Mamba	xLSTM
MIMIC-IV	Dec. 24h	**95.8**	LGBM (Feat.)	eICU	91.5	94.0	95.6	94.1	94.0	93.8	94.5
	Circ. 8h	**94.5**	LGBM (Feat.)	eICU	91.6	92.9	94.5	93.6	93.5	93.3	93.5
	Resp. 24h	**78.5**	LGBM (Feat.)	eICU	72.7	76.2	78.0	76.8	76.5	76.7	76.7
	Kidn. 48h	**89.1**	LGBM (Feat.)	eICU	80.8	84.4	88.0	85.0	83.8	84.7	83.3
eICU	Dec. 24h	**92.3**	LGBM (Feat.)	MIMIC-IV	86.3	90.9	92.2	90.3	90.5	90.5	90.4
	Circ. 8h	95.0	LGBM (Feat.)	MIMIC-IV	92.5	93.9	**95.2**	93.7	93.6	93.7	93.7
	Resp. 24h	**81.4**	LGBM (Feat.)	MIMIC-IV	76.6	79.6	80.4	78.6	78.2	79.2	78.8
	Kidn. 48h	**83.7**	LGBM (Feat.)	MIMIC-IV	71.9	78.7	82.2	77.5	77.4	77.9	77.2
HiRID	Dec. 24h	91.1	LGBM (Feat.)	UMCdb	89.7	92.3	**92.8**	91.9	92.1	92.2	91.8
	Circ. 8h	90.7	LGBM (Feat.)	SICdb	89.7	91.1	**91.5**	90.7	90.7	89.8	90.0
	Resp. 24h	75.3	LGBM (Feat.)	MIMIC-IV	74.1	74.4	75.7	**75.9**	75.8	75.0	74.7
	Kidn. 48h	**93.2**	LGBM (Feat.)	MIMIC-IV	89.9	92.3	**93.2**	92.3	92.0	91.1	90.7
UMCdb	Dec. 24h	89.8	LGBM (Feat.)	HiRID	86.1	90.0	**91.3**	90.6	90.4	89.8	90.1
	Circ. 8h	96.4	GRU	HiRID	95.2	95.9	95.9	96.5	96.3	96.3	**96.6**
	Resp. 24h	**79.7**	LGBM (Feat.)	eICU	77.0	78.1	78.1	79.2	77.7	76.4	76.7
	Kidn. 48h	95.6	LGBM (Last Meas.)	MIMIC-IV	93.1	95.8	**96.2**	95.0	94.4	94.4	93.4
SICdb	Dec. 24h	83.3	LGBM (Feat.)	eICU	79.7	84.3	**84.4**	83.4	83.0	82.3	81.6
	Circ. 8h	89.1	LGBM (Feat.)	HiRID	87.3	88.2	88.8	88.9	**89.4**	88.5	88.9
	Resp. 24h	78.8	LGBM (Feat.)	eICU	76.5	78.2	**79.0**	77.7	77.2	77.6	77.5
	Kidn. 48h	88.8	LGBM (Feat.)	UMCdb	84.2	87.8	**90.0**	85.7	85.9	85.5	85.5
PICdb	Dec. 24h	75.0	LGBM (Feat.)	HiRID	67.9	74.3	**79.6**	66.1	66.0	62.6	61.5
	Circ. 8h	**90.4**	GRU	MIMIC-IV	88.3	89.3	87.5	87.6	87.1	86.4	86.6
	Resp. 24h	**70.8**	GRU	HiRID	66.6	60.6	68.1	66.6	63.0	65.3	65.9
	Kidn. 48h	**71.9**	LGBM (Last Meas.)	HiRID	57.6	67.7	68.5	63.0	61.2	56.7	69.0
Zigong	Dec. 24h	**72.8**	LGBM (Feat.)	UMCdb	66.6	69.3	72.4	69.0	67.0	68.5	68.7
	Circ. 8h	**91.4**	LGBM (Last Meas.)	MIMIC-IV	89.0	88.6	89.0	87.1	86.5	85.2	84.7

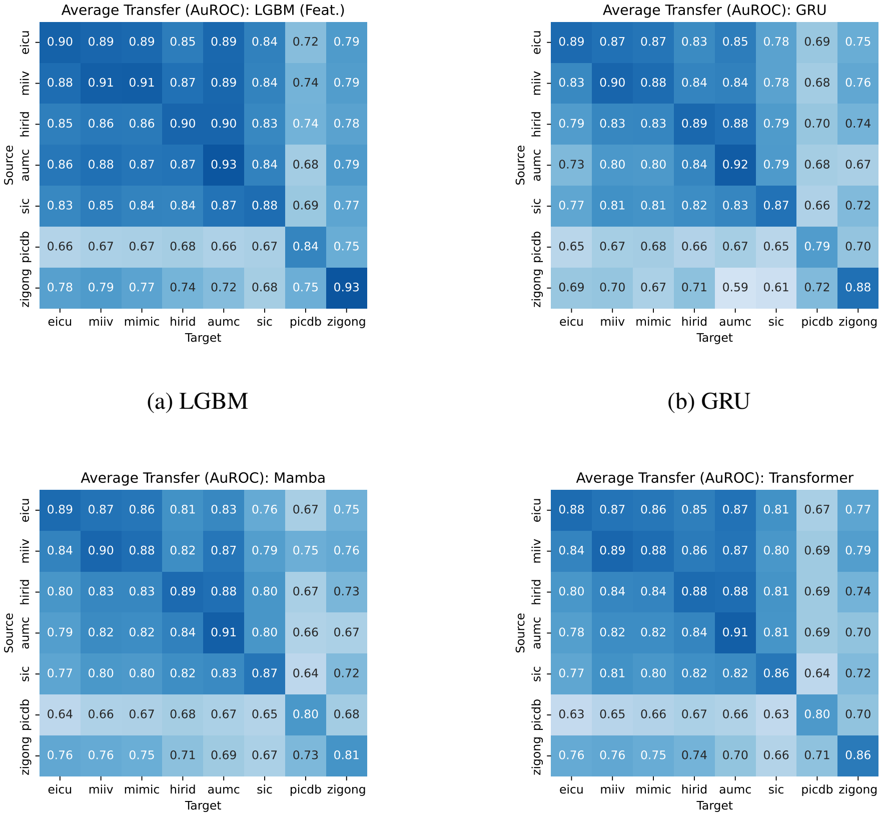
Table 3: Benchmarking results out-of-distribution (AUROC). Bold is best in each row (separately for single-cente and multi-center). Single-center results are an argmax over training datasets while testing on a hold-out dataset. Multi-center models are trained on all but the test dataset. All results show the mean over three different random initialization, except for LR models that are trained using convex optimization.

Dataset	Task	Single-Center			Multi-Center hold-out						
		AUPRC	Best Model	Src. Data	LR (Last Meas.)	LGBM (Last Meas.)	LGBM (Feat.)	GRU	Transformer	Mamba	xLSTM
MIMIC-IV	Dec. 24h	**55.5**	LGBM (Feat.)	eICU	36.9	46.0	54.0	47.6	43.8	48.7	46.8
	Circ. 8h	**62.9**	LGBM (Feat.)	eICU	52.2	56.0	**62.9**	59.0	58.1	57.4	57.9
	Resp. 24h	**78.8**	LGBM (Feat.)	eICU	72.1	76.3	78.4	76.9	76.5	76.9	76.6
	Kidn. 48h	**51.7**	LGBM (Feat.)	eICU	38.9	45.9	48.9	44.1	43.1	44.9	42.5
eICU	Dec. 24h	**38.2**	LGBM (Feat.)	MIMIC-IV	27.1	34.3	37.0	34.3	35.0	33.3	33.7
	Circ. 8h	**62.2**	LGBM (Feat.)	MIMIC-IV	52.9	54.8	61.6	56.4	55.6	57.5	57.5
	Resp. 24h	**77.5**	LGBM (Feat.)	MIMIC-IV	69.5	75.7	76.8	74.4	73.3	74.9	74.2
	Kidn. 48h	**43.8**	LGBM (Feat.)	MIMIC-IV	31.9	39.1	42.3	34.3	34.8	35.3	34.7
HiRID	Dec. 24h	39.3	LGBM (Feat.)	UMCdb	38.4	42.2	42.5	43.1	**45.4**	43.9	42.8
	Circ. 8h	50.6	LGBM (Feat.)	SICdb	49.8	52.8	**53.5**	51.7	52.0	48.3	49.3
	Resp. 24h	89.7	LGBM (Feat.)	MIMIC-IV	87.7	89.2	**90.0**	90.0	89.9	89.4	89.4
	Kidn. 48h	**52.8**	LGBM (Feat.)	MIMIC-IV	39.2	51.4	**55.7**	48.3	47.7	44.0	46.7
UMCdb	Dec. 24h	38.2	LGBM (Feat.)	HiRID	34.3	38.6	**41.9**	41.6	40.9	39.5	39.0
	Circ. 8h	**85.5**	GRU	HiRID	80.0	81.6	83.6	85.4	83.9	84.1	84.9
	Resp. 24h	**86.7**	LGBM (Feat.)	eICU	83.8	85.5	86.6	86.3	85.1	84.4	84.6
	Kidn. 48h	59.4	LGBM (Last Meas.)	MIMIC-IV	46.7	56.9	**60.6**	52.6	49.9	50.9	45.4
SICdb	Dec. 24h	28.6	LGBM (Feat.)	eICU	24.1	**31.0**	30.8	29.4	28.0	28.3	28.9
	Circ. 8h	46.9	LGBM (Feat.)	HiRID	46.1	47.9	48.1	46.2	**48.7**	45.8	46.3
	Resp. 24h	87.6	LGBM (Feat.)	eICU	84.8	87.0	**87.7**	86.9	86.6	86.6	86.7
	Kidn. 48h	37.0	LGBM (Feat.)	UMCdb	25.4	36.0	**41.6**	31.9	32.1	32.1	30.9
PICdb	Dec. 24h	7.9	LGBM (Feat.)	HiRID	5.9	7.3	**8.3**	6.5	4.5	5.1	2.7
	Circ. 8h	**89.4**	GRU	MIMIC-IV	86.7	86.9	85.6	86.8	86.1	85.6	85.4
	Resp. 24h	**7.7**	GRU	HiRID	5.1	4.1	7.3	5.8	5.8	5.2	4.8
	Kidn. 48h	8.1	LGBM (Last Meas.)	HiRID	6.6	7.3	7.5	7.7	8.0	7.9	**8.8**
Zigong	Dec. 24h	**22.2**	LGBM (Feat.)	UMCdb	20.8	19.1	20.3	17.2	18.1	18.9	18.8
	Circ. 8h	**98.6**	LGBM (Last Meas.)	MIMIC-IV	98.2	97.9	98.2	97.9	97.7	97.0	97.1

Table 4: Benchmarking results out-of-distribution (AUPRC). Bold is best in each row (separately for single-cente and multi-center). Single-center results are an argmax over training datasets while testing on a hold-out dataset. Multi-center models are trained on all but the test dataset. All results show the mean over three different random initialization, except for LR models that are trained using convex optimization.

A.3 Transfer Heatmaps

Figure 4 shows single-center task-averaged transfer results for LGBM [28], GRU [7], Mamba [9] and Transformer [50].



(a) LGBM

(b) GRU

(c) Mamba

(d) Transformer

Figure 4: Single-center task-averaged transfer performance heatmaps (AUROC).

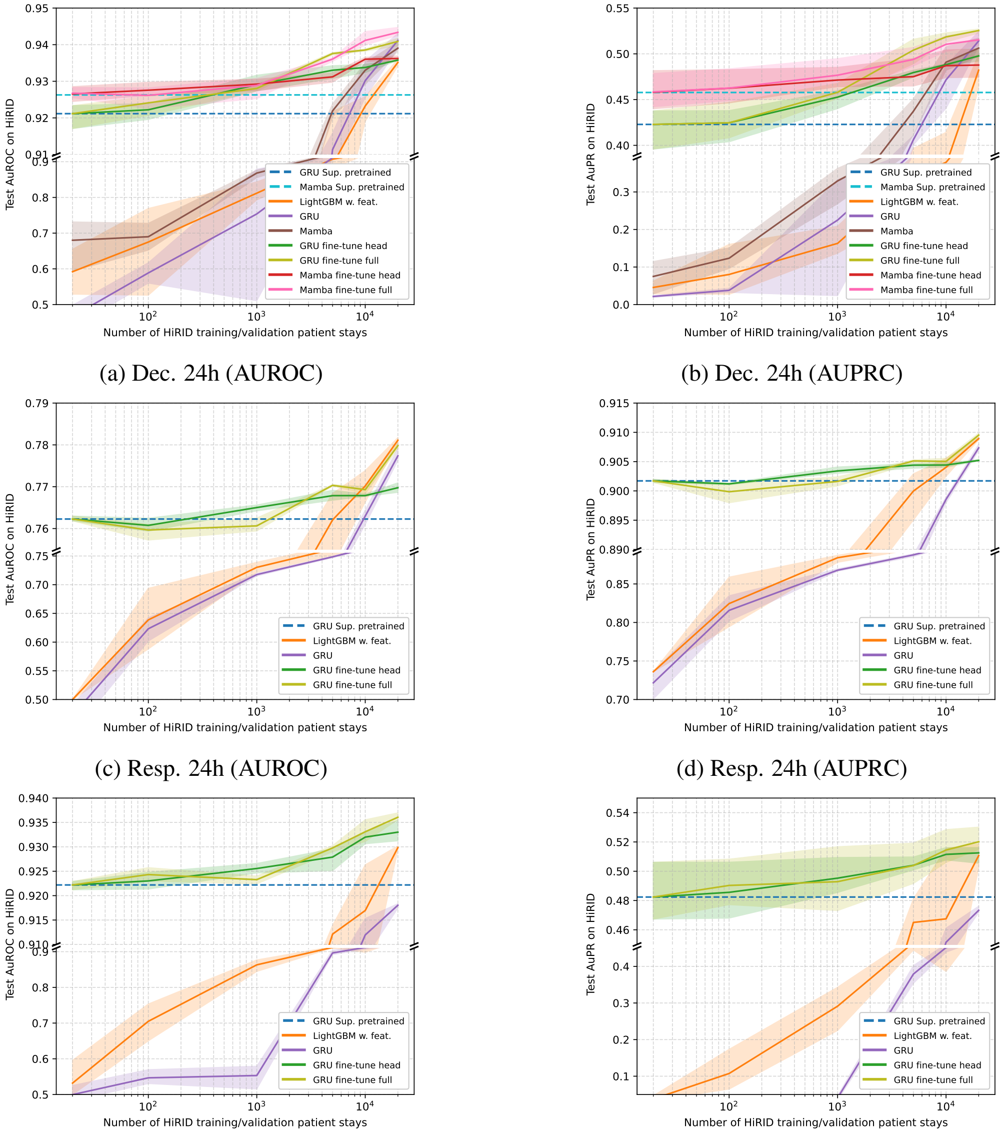
A.4 Disposition Prediction

Models	Disposition
LR (Last Meas.)	72.7
LR	78.2
LGBM (Last Meas.)	74.8 ± 0.02
LGBM (Feat.)	$\mathbf{80.4 \pm 0.01}$
GRU	79.9 ± 0.04
Transformer	79.9 ± 0.03
Mamba	79.9 ± 0.09
xLSTM	80.1 ± 0.16

Table 5: Disposition prediction on MIMIC-IV-ED (AUROC).

A.5 Fine-tuning Study

In Figure 5 we show further fine-tuning study results for respiratory failure and kidney failure predictions, which confirm the trend already highlighted and discussed in Figure 3. The performed data harmonization work is highly valuable for small to medium-sized hospitals, which only have limited amounts of training patients available and can thus significantly benefit from pretraining on data from other hospitals.



(a) Dec. 24h (AUROC)

(b) Dec. 24h (AUPRC)

(c) Resp. 24h (AUROC)

(d) Resp. 24h (AUPRC)

(e) Kidney 24h (AUROC)

(f) Kidney 24h (AUPRC)

Figure 5: Supervised fine-tuning study performed on HiRID for decompensation (Figures 5a and 5b, respiratory failure (Figures 5d and 5e), and kidney failure (Figures 5e and 5f) by progressively increasing the number of patients shown during training or fine-tuning. *GRU* and *LGBM w. feat.* are trained from scratch using HiRID data only. *GRU pretrained* is trained on all data excluding HiRID patients. *GRU fine-tuned (head/full)* initialize the network with *GRU pretrained* and fine-tune the full network or only the single linear logit head.

	Datasets									Features		Transfer		
	MIMIC-III	eICU	UMCdb	HiRID	MIMIC-IV	SICdb	PICdb	Zigong EHR	MIMIC-IV-ED	Harm. treatments	Multi-unit	Single-center	Multi-center	Fine-tuning
[45]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
[34]	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
[48]	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 6: Multi-dataset critical care time-series benchmarks

B Related Work

Time Series Foundation Models Following advances in natural language processing (NLP), Large-scale multi-purpose pretrained models referred to as "foundation models" have sprouted across data and application types. When considering time series, a large body of works has exclusively focused on forecasting non-medical data [10, 16, 60, 1].

Foundation Models for Healthcare In the clinical domain, existing foundation models have not considered in-hospital critical care time series but rather other forms of Electronic Health Records (EHR) such as billing codes [54] and medical reports [6]. Some studies [18, 55] explored the adaptability of public EHR models for clinical prediction tasks, while others [17] evaluated their effectiveness in improving in-distribution and out-of-distribution performance.

Benchmarks on ICU time-series In the literature, we observe two benchmarking strategies: single- and multi-center. Harutyunyan et al. [19] provided the first standardized and reproducible single-center benchmarks built on MIMIC-III [24]. Following this seminal work, a line of studies emerged that defined new tasks or explored different datasets, such as Sheikhalishahi et al. [42] on eICU [38], Yèche et al. [58] on HiRID [12], and Wang et al. [52] as an alternative on MIMIC-III. The proliferation of work around single-center data has led researchers to aggregate them into multi-center studies such as [34] and [48]. We present a comparison in Table 6. It is important to emphasize that, unlike previous efforts, we both perform a new largest to-date dataset harmonization and build a comprehensive benchmark.

C Data Harmonization and Processing

C.1 Data Sources

The overview of the datasets is shown in Table 7. We do not harmonize RICD [51] as it is not free access (a separate contract and payment are required). Other datasets are available via PhysioNet [14] or directly from the providers.

To the best of our knowledge, this is the first work bringing together critical care datasets from the US, Europe, and, for the first time, China. Harmonizing datasets across different continents can improve generalization and is crucial to the fairness and inclusiveness of ML research on critical care data. The diversity of our dataset enables research for small but specific cohorts of patients. For example, PICdb [30] is a small pediatric dataset. The average age is under one year, while it is over 60 on other datasets (see Table 7 in Appendix C). By providing an easy way to pretrain on large amounts of data, we create an opportunity for smaller-scale targeted studies to benefit from the existing larger-scale research on modeling for critical care time series.

[11] Certificate "Data or Specimens Only Research" from the Collaborative Institutional Training Initiative (CITI) program: `physionet.org/about/citi-course/`

		MIMIC-III	MIMIC-IV	eICU	UMCdb§	HiRID	SICdb	PICdb	Zigong EHR	MIMIC-IV-ED	~~RICD~~
General†	Time	2001-12	2008-19	2014-15	2003-16	2008-16	2013-21	2010-18	2019-20	2011-19	2017-23
	Country	US	US	US	NL	CH	AU	CN	CN	US	RU
	Easy access*	+	+	+	±	±	±	+	+	+	−
	Max resolution, min	60	60	60	60	2	1	5	60	–	6
	Admissions	61532	76540	200859	23106	33905	27386	13499	2790	∼ 425000	3291
	Patients	46476	53150	–	20109	–	–	12881	2790	–	2562
	Mean LoS, days	2.1	11	1.57	1.08	0.95	3.5	9.3	4	–	32
	Mean age, years	65.8	64.7	65	65	65	–	0.8	69.2	–	57.8
	Mortality, %	8.5	11.6	9.94	12.05	6.52	3.45	6.9	5.77	–	12.31
Extracted	Resolution, min	60	60	60	5	5	5	60	60	60	–
	Admissions	53713	70831	183695	22889	33558	24522	13295	2525	177714	–
Label, %	Decompensation 24h	8	7	5	10	6	5	7	43	–	–
	Circulatory 8h	14	19	7	23	31	30	16	32	–	–
	Respiratory 24h	22	25	14	45	45	54	2	0	–	–
	Kidney 48h	5	7	7	5	3	5	1	0	–	–
	ED Disposition	–	–	–	–	–	–	–	–	51	–

Table 7: Datasets overview.

* Denoted as + if only a CITI certificate[11]and ± if additional provider approval is required, − otherwise.
§ Shortened AmsterdamUMCdb to "UMCdb".
† We consider ED admissions for MIMIC-IV-ED, and ICU admissions for other datasets.

By incorporating both ICU and ED datasets, we provide a way to study joint ED-ICU models, potentially leading to a unified clinical prediction model regardless of the hospital unit.

C.2 Inclusion criteria

We consider patient stays that after extraction have a valid admission and discharge time, a valid length of stay (LoS) that is longer than 4 fours, a maxium gap between measurements smaller than 48 hour, and more than 4 measurements.

Compared to Van De Water et al. [48], we broaden the inclusion criteria by reducing the LoS requirement from 6 to 4 hours, and increasing the allowed maximum gap between measurements from 12 to 48 hours.

By including as many patients as possible, we aim to create a more general version of the dataset, that can be further trimmed down for specific studies. Additionally, a wide range of stays can improve the generalizability of predictive models.

C.3 Harmonization

The datasets we consider are recorded using different, non-standardized, formats. We perform dataset harmonization with the `ricu` package as a basis Bennett et al. [4]. `ricu` defines data source agnostic concepts as an abstraction for encoding clinical concepts. These include static information about the patient (e.g., height), observations (e.g., heart rate), and treatments (e.g., administration of antibiotics). By mapping the concepts to source variables from each dataset the package facilitates exporting a unified view of the data across all of them.

Expanding prior work [48, 33, 4], we implement new observation concepts and incorporate new ICU datasets, namely SICdb, PICdb, Zigong EHR, creating the largest harmonized ICU dataset to date. Further, we integrate an ED dataset (MIMIC-IV-ED), increasing the number of processed stays from around 400,000 [48] to over 600,000. The expansion increases the total number of final extracted individual data points from approximately 400 million close to one billion.

Crucially, we introduce a principled way to harmonize a wide range of treatment variables. This significantly increases the number of concepts compared to previous works [33, 48]. We define the new concepts using clinical expert opinion informed by what variables were reported as most important for various tasks and models in the literature (see Table 8, Appendix C). Previous works [33] have suggested that including medication variables harms the accuracy of predictive models,

Variables				Clinical importance		Task importance[*]				Literature Importance[†]			
Meta variable	Type	Organ system	Group	Included as	Priority	Circ.	Resp.	Kidn.	Sepsis	CircEWS	RMS	KDIGO	Moor
Dobutamine	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✓	✓	✓	✓	✓	✗	✗	✓
Levosimendan	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✓	✗	✓	✗	✓	✗	✓	✗
Norepinephrine	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✗	✓	✗	✓	✗	✓	✗	✓
Epinephrine	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✗	✓	✗	✓	✗	✓	✗	✓
Milrinone	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✓	✗	✗	✗	✓	✗	✗	✗
Theophylline	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✓	✗	✗	✗	✓	✗	✗	✗
Dopamine	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✗	✗	✗	✓	✗	✗	✗	✓
Vasopressin	Drug	Cardiovascular	Vasopressor / Inotropes	Rate & ind.	High	✗	✗	✗	✗	✗	✗	✗	✗
Heparin	Drug	Cardiovascular	Anticoagulants	Rate & ind.	High	✗	✓	✓	✗	✗	✓	✓	✗
Propofol	Drug	Nervous	Sedatives / Anxiolytics	Rate & ind.	High	✗	✓	✗	✗	✗	✓	✗	✗
Benzodiacepine	Drug	Nervous	Sedatives / Anxiolytics	Rate & ind.	High	✗	✓	✗	✗	✗	✓	✗	✗
Loop diuretic	Drug	Renal	Diuretics	Rate & ind.	High	✗	✓	✓	✗	✗	✓	✓	✗
Other sedatives	Drug	Nervous	Sedatives / Anxiolytics	Indicator	High	✗	✗	✗	✗	✗	✗	✗	✗
Opiate painkillers	Drug	Nervous	Pain killers	Indicator	High	✗	✗	✓	✗	✗	✗	✓	✗
Non-opioid analgesic	Drug	Nervous	Pain killers	Indicator	High	✓	✗	✗	✗	✓	✗	✗	✗
Paralytics	Drug	Nervous	Paralyzing	Indicator	High	✗	✗	✗	✗	✗	✗	✗	✗
Administration of antibotics	Drug	Infectious	Antibiotics	Indicator	High	✗	✗	✓	✗	✗	✗	✓	✗
Insulin	Drug	Endocrine	Insulin	Indicator	High	✗	✓	✗	✗	✗	✓	✗	✗
Anti delirant medi	Drug	Nervous	Anti delirant medi	Indicator	Med	✗	✗	✓	✗	✗	✗	✓	✗
Other diuretics	Drug	Renal	Diuretics	Indicator	Med	✗	✗	✗	✗	✗	✗	✗	✗
Other anticoagulants	Drug	Cardiovascular	Anticoagulants	Indicator	Med	✗	✗	✗	✗	✗	✗	✗	✗
Vasodilators	Drug	Cardiovascular	Antihypertensive + Vasodilators	Indicator	Med	✗	✗	✗	✗	✗	✗	✗	✗
Antiarrhythmics	Drug	Cardiovascular	Antiarrhythmic	Indicator	Med	✗	✗	✗	✗	✗	✗	✗	✗
Packed red blood cells	Blood	Cardiovascular / Renal	Infusion of blood products	Indicator	Med	✗	✗	✓	✗	✗	✗	✓	✗
FFp	Blood	Cardiovascular / Renal	Infusion of blood products	Indicator	Med	✗	✗	✗	✗	✗	✗	✗	✗
Platelets	Blood	Cardiovascular / Renal	Infusion of blood products	Indicator	Med	✗	✗	✗	✗	✗	✗	✗	✗
Albumin	Blood	Cardiovascular / Renal	Infusion of blood products	Indicator	Med	✗	✗	✗	✗	✗	✗	✗	✗
Fluid administration	Feeding / Electrolyte	Gastrointestinal / Renal	Electrolytes	Indicator	Med	✗	✗	✓	✗	✗	✓	✗	✗
Electrolytes-Phosphate	Feeding / Electrolyte	Gastrointestinal / Renal	Electrolytes	None	Low	✗	✗	✗	✗	✗	✗	✗	✗
Electrolytes-Kalium	Feeding / Electrolyte	Gastrointestinal / Renal	Electrolytes	None	Low	✗	✗	✓	✗	✗	✗	✗	✗
Elektrolytes-Mg	Feeding / Electrolyte	Gastrointestinal / Renal	Electrolytes	None	Low	✗	✗	✓	✗	✗	✗	✓	✗
Enteral feeding	Feeding / Electrolyte	Gastrointestinal / Renal	Feeding	None	Low	✗	✗	✓	✗	✗	✗	✓	✗
Parenteral feeding	Feeding / Electrolyte	Gastrointestinal / Renal	Feeding	None	Low	✗	✗	✓	✗	✗	✗	✓	✗
Glucose	Drug	Endocrine	Glucose	None	Low	✗	✗	✗	✗	✗	✗	✗	✗
Antiepileptic	Drug	Nervous	Antiepileptic	None	Low	✗	✗	✗	✗	✗	✗	✗	✗
Inhalation	Drug	Respiratory	Inhalation	None	Low	✗	✗	✗	✗	✗	✗	✗	✗
Platelet inhibitors	Drug	Cardiovascular	Platelet inhibitors	None	Low	✗	✗	✓	✗	✗	✗	✓	✗
Desmopressin	Drug	Cardiovascular	Vasopressor / Inotropes	None	Low	✗	✗	✗	✗	✗	✗	✗	✗
Inhalation	Drug	Respiratory	Inhalation	None	Low	✗	✗	✗	✗	✗	✗	✗	✗
Immunmodulation	Drug	Immune	Immunmodulation	None	Low	✗	✗	✗	✗	✗	✗	✗	✗
Laxatives	Drug	Gastrointestinal	Laxatives	None	Low	✗	✗	✓	✗	✗	✗	✓	✗
Peritoneal dialysis	Blood	Cardiovascular / Renal	Dialysis	None	Low	✗	✗	✓	✗	✗	✗	✓	✗
Infusion of blood products	Blood	Cardiovascular / Renal	Blood products	None	Low	✗	✗	✓	✗	✗	✗	✗	✗
Supplemental oxygen	Ventilator	Respiratory	Respirator settings	Rate	Med	✓	✓	✗	✗	✓	✓	✗	✗
Other ventilator settings	Ventilator	Respiratory	Respirator settings	None	Low	✗	✓	✓	✓	✗	✓	✗	✓

Table 8: Treatment concepts and their importance in clinical practice and ML literature.

[*] Circ. is circulatory, Resp. is respiratory, and Kidn. is kidney failure.

[†] CircEWS [22], RMS [21], KDIGO [32], Moor [34].

but little research has been done into the reasons behind this effect and what can be done to mitigate it. The information about administered medications is an insight into the actions of the clinicians, and could drastically improve the model accuracy and transfer. By including these variables in the harmonization pipeline, we prepared the ground for deeper investigation in this direction.

Oliver et al. [36] have proposed a processing pipeline for a subset of the source datasets considered in this work and harmonized treatments by including drug exposure information as indicators. We improve on this by (1) considering not only indicators, but also administration rates for core medications used in critical care settings, and (2) grouping individual drugs into abstract treatment concepts, thereby increasing the overlap across datasets in concepts while maintaining relevance for downstream applications.

Ultimately, providing harmonized treatment information including administered dosages across a collection of datasets enables future research on learning generalizable treatment effect estimations on critical care time series.

C.3.1 Concepts for treatment variables

In Table 8 we present concepts for treatments that were identified as important in the ML literature with their clinical importance. The choice of concepts balances granularity, missingness, and time effort, to incorporate as much of the signal from the data as possible while keeping missingness across datasets low and the variable labeling feasible for the medical experts.

The statistics for the new concepts covering medications are shown in Table 9. We note that we include all medications as indicators, and, for the most important ones, rates if possible to compute (e.g., the information on dosage is included with a convertible unit and appropriate time information is available).

	Concept name	Abbreviation	MIMIC-IV	HiRID	MIMIC-III	PICdb	SICdb	Zigong EHR	eICU	UMCdb	MIMIC-IV-ED	Total
Rates & indicators	dobutamine	dobu	1	1	4	1	3	1	13	1	2	27
	levosimendan	levo	0	1	0	0	1	1	0	0	0	3
	norepinephrine	norepi	1	5	3	1	5	1	41	1	6	64
	epinephrine	epi	2	5	8	1	1	1	20	1	10	49
	milrinone	milirin	1	2	2	2	1	1	10	0	2	21
	theophylline	teophyllin	1	6	3	2	1	2	6	4	0	25
	dopamine	dopa	1	0	7	1	1	1	17	1	2	31
	vasopressin	adh	1	2	17	0	1	0	22	0	3	46
	heparin	hep	4	3	52	1	1	2	75	2	8	148
	propofol	prop	2	13	4	1	5	3	39	1	3	71
	benzodiacepine	benzdia	3	18	19	7	9	7	110	13	45	231
	loop diuretic	loop_diur	3	8	9	2	4	4	62	2	7	101
Indicators	other sedatives	sed	9	5	36	3	16	7	57	8	24	165
	opiate painkiller	op_pain	7	32	52	9	21	12	309	13	86	541
	non-opioid analgesic	nonop_pain	2	32	9	14	12	14	163	11	38	295
	paralytic	paral	7	0	56	2	3	4	100	5	7	184
	antibotics	abx	55	125	208	164	49	90	330	53	99	1173
	insulin	ins	8	5	20	6	7	13	132	7	6	204
	fluid administration	fluid	7	2	59	0	23	14	297	9	7	418
	packed red blood cells	inf_rbc	3	2	3	0	0	0	14	1	0	23
	fresh frozen plasma	ffp	3	2	8	0	0	0	7	1	0	21
	platelets	plat	3	2	3	0	1	0	7	1	0	17
	albumin infusion	inf_alb	2	0	14	1	4	2	57	1	0	81
	anti deliriant	anti_delir	1	16	0	1	0	2	9	1	4	34
	other diuretics	oth_diur	1	10	28	3	13	1	46	6	10	118
	other anticoagulants	anti_coag	8	18	43	6	15	16	161	14	20	301
	antihypertensive and vasodilators	vasod	11	78	62	14	65	53	313	41	84	721
	antiarrhythmic	anti_arrhythm	14	8	23	8	14	8	95	11	27	208
	Used		442	891	9458	719	1595	1077	9763	1113	1085	26143
	Not used		292	492	8751	470	1332	818	7278	908	602	20943
	Total		453	893	9503	720	1608	1078	9790	1117	1102	26264

Table 9: Presence of medication concepts across datasets

We labeled but did not include the data in the MIMIC-III prescriptions table because it only specifies the prescription and not the drug administration.

C.3.2 Concept reference table

The full concept (or variable) reference table is shown in Table 10. It includes 141 variable: 6 static demographic, 80 observation, and 55 treatment variables.

Tag	Name	Type	Organ System	Unit
map	Mean Arterial Blood Pressure	observation	circulatory	mmHg
lact	Lactate	observation	circulatory	mmol/L
age	Age	demographic	None	years
weight	Weight	demographic	None	kg
sex	Sex	demographic	None	categorical
height	Height	demographic	None	cm
hr	Heart Rate	observation	circulatory	bpm
fio2	FiO2	observation	respiratory	%
resp	Respiratory Rate	observation	respiratory	insp/min
temp	Temperature	observation	infection	C
crea	Creatinine	observation	metabolic_renal	mg/dL
urine_rate	Urine Rate Per Hour	observation	metabolic_renal	mL/h
po2	Partial Pressure Of Oxygen	observation	respiratory	mmHg
ethnic	Ethnic Group	demographic	None	categorical
alb	Albumin	observation	gastrointestinal	g/dL
alp	Alkaline Phosphatase	observation	gastrointestinal	IU/L
alt	Alanine Aminotransferase	observation	gastrointestinal	IU/L
ast	Aspartate Aminotransferase	observation	gastrointestinal	IU/L
be	Base Excess	observation	metabolic_renal	mmol/l
bicar	Bicarbonate	observation	metabolic_renal	mmol/l
bili	Total Bilirubin	observation	gastrointestinal	mg/dL
bili_dir	Bilirubin Direct	observation	gastrointestinal	mg/dL
bnd	Band Form Neutrophils	observation	infection	%
bun	Blood Urea Nitrogen	observation	metabolic_renal	mg/dL
ca	Calcium	observation	metabolic_renal	mg/dL
cai	Calcium Ionized	observation	metabolic_renal	mmol/L
ck	Creatine Kinase	observation	circulatory	IU/L
ckmb	Creatine Kinase MB	observation	circulatory	ng/mL
cl	Chloride	observation	metabolic_renal	mmol/l
crp	C-Reactive Protein	observation	infection	mg/L
dbp	Diastolic Blood Pressure	observation	circulatory	mmHg
fgn	Fibrinogen	observation	circulatory	mg/dL
glu	Glucose	observation	metabolic_renal	mg/dL
hgb	Hemoglobin	observation	circulatory	g/dL
inr_pt	Prothrombin	observation	circulatory	INR
k	Potassium	observation	metabolic_renal	mmol/l
lymph	Lymphocytes	observation	infection	%
methb	Methemoglobin	observation	circulatory	%
mg	Magnesium	observation	metabolic_renal	mg/dL
na	Sodium	observation	metabolic_renal	mmol/l
neut	Neutrophils	observation	infection	%
pco2	CO2 Partial Pressure	observation	respiratory	mmHg
ph	pH Of Blood	observation	metabolic_renal	pH
phos	Phosphate	observation	metabolic_renal	mg/dL
plt	Platelet Count	observation	circulatory	G/l
ptt	Partial Thromboplastin Time	observation	circulatory	sec
sbp	Systolic Blood Pressure	observation	circulatory	mmHg
tnt	Troponin T	observation	circulatory	ng/mL
wbc	White Blood Cell Count	observation	infection	G/l
basos	Basophils	observation	infection	%
eos	Eosinophils	observation	infection	%
mgcs	Glasgow Comma Scale Motor	observation	neuro	categorical
tgcs	Glasgow Comma Scale Total	observation	neuro	categorical
vgcs	Glasgow Comma Scale Verbal	observation	neuro	categorical
egcs	Glasgow Comma Scale Eye	observation	neuro	categorical
hct	Hematocrit	observation	circulatory	%
rbc	Red Blood Cell Count	observation	circulatory	m/uL
tri	Troponin I	observation	circulatory	ng/mL
etco2	Endtital CO2	observation	respiratory	mmHg
rass	Richmond Agitation Sedation Scale	observation	neuro	categorical
hbco	Carboxyhemoglobin	observation	circulatory	%
esr	Erythrocyte Sedimentation Rate	observation	infection	mm/hr
pt	Prothrombine Time	observation	circulatory	sec
adm	Patient Admission Type	demographic	None	categorical
hba1c	Hemoglobin A1C	observation	metabolic_renal	%
samp	Body Fluid Sampling, Detected Bacterial Growth	observation	infection	categorical
spo2	Pulse Oxymetry Oxygen Saturation	observation	respiratory	%
sao2	Oxygen Saturation In Arterial Blood	observation	respiratory	%
icp	Intra Cranial Pressure	observation	neuro	mmHg
cout	Cardiac Output	observation	circulatory	l/min
mpap	Mean Pulmonal Arterial Pressure	observation	circulatory	mmHg
spap	Systolic Pulmonal Arterial Pressure	observation	circulatory	mmHg

dpap	Diastolic Pulmonal Arterial Pressure	observation	circulatory	mmHg
cvp	Central Venous Pressure	observation	circulatory	mmHg
svo2	Mixed Venous Oxygenation	observation	circulatory	%
pcwp	Pulmonary Capillary Wedge Pressure	observation	circulatory	mmHg
peep	Positive End Expiratory Pressure - Mechanical Ventilation	observation	respiratory	cmH2O
peak	Peak Pressure - Mechanical Ventilation	observation	respiratory	cmH2O
plateau	Plateau Pressure - Mechanical Ventilation	observation	respiratory	cmH2O
ps	Pressure Support - Mechanical Ventilation	observation	respiratory	cmH2O
tv	Tidal Volume	observation	respiratory	ml
airway	Type Of Airway Ventilation	observation	respiratory	categorical
supp_o2_vent	Supplemental Oxygen From Ventilator	treatment	respiratory	%
ygt	Gamma GT	observation	gastrointestinal	U/L
amm	Ammoniak	observation	gastrointestinal	mmol/L
amyl	Amylase	observation	gastrointestinal	U/L
lip	Lipase	observation	gastrointestinal	U/L
ufilt	Ultrafiltration On Continuous RRT	treatment	metabolic_renal	ml
ufilt_ind	Ultrafiltration On Continuous RRT Indicator	treatment	metabolic_renal	indicator
dobu	Dobutamine	treatment	circulatory	mcg/min
levo	Levosimendan	treatment	circulatory	mcg/min
norepi	Norepinephrine	treatment	circulatory	mcg/min
epi	Epinephrine	treatment	circulatory	mcg/min
milrin	Milrinone	treatment	circulatory	mcg/min
teophyllin	Theophylline	treatment	circulatory	mg/min
dopa	Dopamine	treatment	circulatory	mcg/min
adh	Vasopressin	treatment	circulatory	U/min
hep	Heparin	treatment	circulatory	U/h
prop	Propofol	treatment	neuro	mcg/min
benzdia	Benzodiacepine	treatment	neuro	mg/h
sed	Other Sedatives	treatment	neuro	indicator
op_pain	Opiate Painkiller	treatment	neuro	indicator
nonop_pain	Non-Opioid Analgesic	treatment	neuro	indicator
paral	Paralytic	treatment	neuro	indicator
abx	Antibotics	treatment	infection	indicator
loop_diur	Loop Diuretic	treatment	metabolic_renal	mg/h
ins_ind	Insulin	treatment	None	indicator
fluid	Fluid Administration	treatment	None	indicator
inf_rbc	Packed Red Blood Cells	treatment	None	indicator
ffp	Fresh Frozen Plasma	treatment	None	indicator
plat	Platelets	treatment	None	indicator
inf_alb	Albumin Infusion	treatment	None	indicator
anti_delir	Anti Deliriant	treatment	neuro	indicator
oth_diur	Other Diuretics	treatment	metabolic_renal	indicator
anti_coag	Other Anticoagulants	treatment	circulatory	indicator
vasod	Antihypertensive And Vasodilators	treatment	circulatory	indicator
anti_arrhythm	Antiarrhythmic	treatment	circulatory	indicator
dobu_ind	Dobutamine Indicator	treatment	circulatory	indicator
levo_ind	Levosimendan Indicator	treatment	circulatory	indicator
norepi_ind	Norepinephrine Indicator	treatment	circulatory	indicator
epi_ind	Epinephrine	treatment	circulatory	indicator
milrin_ind	Milrinone Indicator	treatment	circulatory	indicator
teophyllin_ind	Theophylline Indicator	treatment	circulatory	indicator
dopa_ind	Dopamine Indicator	treatment	circulatory	indicator
adh_ind	Vasopressin Indicator	treatment	circulatory	indicator
hep_ind	Heparin Indicator	treatment	circulatory	indicator
prop_ind	Propofol Indicator	treatment	circulatory	indicator
benzdia_ind	Benzodiacepine Indicator	treatment	circulatory	indicator
loop_diur_ind	Loop Diuretics Indicator	treatment	circulatory	indicator
dobu_ind	Dobutamine Indicator	treatment	circulatory	indicator
levo_ind	Levosimendan Indicator	treatment	circulatory	indicator
norepi_ind	Norepinephrine Indicator	treatment	circulatory	indicator
epi_ind	Epinephrine Indicator	treatment	circulatory	indicator
milrin_ind	Milrinone Indicator	treatment	circulatory	indicator
teophyllin_ind	Theophylline Indicator	treatment	circulatory	indicator
dopa_ind	Dopamine Indicator	treatment	circulatory	indicator
adh_ind	Vasopressin Indicator	treatment	circulatory	indicator
hep_ind	Heparin Indicator	treatment	circulatory	indicator
prop_ind	Propofol Indicator	treatment	neuro	indicator
benzdia_ind	Benzodiacepine Indicator	treatment	neuro	indicator
loop_diur_ind	Loop Diuretic Indicator	treatment	metabolic_renal	indicator

Table 10: Concept reference

C.4 Processing

We use anonymized data with permissive exclusion criteria (Appendix C.2) to include as many patients as possible. The time series are extracted as a uniform grid at resolutions of 5 and 60 minutes depending on the dataset balancing sampling precision and interoperability (see Table 7, Appendix C). Further, similar to Yèche et al. [58], we remove outliers, impute missing values, scale variables, extract features for tree-based models, and define task labels.

Finally, we export the processed data into two formats consumable by modern deep learning and classical machine learning algorithms. First, a dense fully imputed time-grid (including feature extraction if applicable for the model) and second a tokenized data format [15, 20], which encodes only ground truth measured data points as a triplet of time, variable, and observed value. The second format removes the need for imputation and has recently been proposed as a more suitable data representation format for scaling models on highly irregular time-series data [47] and sharing of data processing outputs [2]. Further details on data harmonization and processing are described in Appendix C.

C.4.1 Data Scaling

Data is scaled depending on its type:

- continuous observations are standardized (i.e. centered and scaled to unit variance),
- categorical observations are one-hot encoded and each variable has a dedicated class to encode missing information,
- continuous treatments are quantile-transformed and mapped to the $[0, 1]$ range such that a 0 represents *no medication given*,
- treatment indicators are binary encoded using $\{0, 1\}$.

C.4.2 Imputation

Gridded time-step data as inputs for model training are forward-filled indefinitely for all observation variables. The remaining missing values are then imputed with 0 for continuous variables, which corresponds to a population mean imputation after considering standard scaling before the imputation stage. The remaining categorical entries are imputed with a value corresponding to the dedicated class that encodes missing information for each categorical variable.

Any treatment variable is excluded from forward-filling operations and missing data points are strictly filled using 0, which given the previously introduced scaling and encoding scheme always corresponds to no treatment being applied.

C.4.3 Feature Extraction

We build on the feature set proposed by Soenksen et al. [44] to process the MIMIC-IV [23] dataset. To improve performance we then further expand this set of features and select specific features for each variable type. For each time-step, each feature is computed over three history sizes of 8, 24, and 72 hours:

- For continuous observations and continuous treatment variables, we compute:
  - mean on raw and imputed data,
  - standard deviation on raw data,
  - slope of a linear fit on the raw data and imputed data,
  - mean absolute change over imputed data,
  - fraction of non-missing data points,
  - quantiles: 0% (Min.), 10%, 50%, 90%, 100% (Max.).
- For categorical variables, we compute the mode, number of missing points, and a binary indicator of whether there are any missing points at all.
- For treatment indicators we compute the number of points with treatment and a binary indicator whether any treatment was applied.

C.4.4 Task Annotations

Our study focuses on clinically relevant real-time prediction tasks where patient outcomes in ICU can be influenced by timely intervention. These include: circulatory failure [22], respiratory failure [21], and kidney function [32]. Additionally, to ensure a diverse range of tasks and to facilitate comparison with previous works, we include the prediction of decompensation [19]. All of these are modeled as binary early event prediction tasks [59] with a clinically relevant prediction horizon. On the emergency department data, we consider disposition prediction [5, 29].

We define sample labels by annotating the time series following the clinical definitions. Most importantly, we annotate a positive and a negative case only if there's enough evidence in favor of either. Based on these cases, we define early event prediction labels (e.g., respiratory failure) that are then used for online classification of the future state of the patient.

Early event prediction (EEP) label for a given time step is computed as follows: (1) a detection (positive EEP label) is marked if any time-point in the future within the horizon is annotated as the patient is in a failure state; (2) a negative EEP label (a stable patient without any upcoming failure state) is annotated only if there is no failure state annotation and there is at least one confirmed stable state within the horizon; (3) if there is no data confirmed evidence for either the patient being in failure or being stable within the horizon, no EEP label is assigned and no training and evaluation is performed for that specific time step. We use task-specific and clinically relevant prediction horizons from existing literature (8 hours for circulatory failure [22], 24 hours for decompensation [19] and respiratory failure [21], and 48 hours for kidney failure [32]).

D Training details

We trained deep learning models using AdamW optimizer [31], with a cross-entropy objective for classification tasks. We evaluated models using task-specific metrics: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) for classification tasks. For all models, we tuned a subset of important hyper-parameters using grid search. Each set of parameters was run with 3 different random initializations and we report mean metric performance (standard deviations are shown in tables if space permits it). Further details are presented in Appendix D.

In the fine-tuning study, *GRU* and *LGBM w. feat.* are trained from scratch using HiRID data only. *GRU pretrained* is trained on all data excluding HiRID patients. *GRU fine-tuned (head/full)* initialize the network with *GRU pretrained* and fine-tune the full network or only the single linear logit head.

Deep learning approaches were implemented in `pytorch` [37]. All metrics were computed using `torchmetrics` [35].

We evaluate performance of 7 model architectures. For each, we find the best set of hyperparameters using grid search with a set of approximately 12 points per model. For deep learning architectures we focus on hidden dimensions, number of layers, and architecture-specific parameters. For LightGBM [28] we choose a strong starting point based on hyperparameters reported by Yèche et al. [58], Hyland et al. [22] and then further tune: `colsample_bytree`, `subsample`, num_leaves, `min_child_samples`, and `subsample_for_bin`. For linear models trained using `glum` [39] we optimize regularization parameters. For each set of hyperparameters, model performance is evaluated as an average across three seeds.

Single center experiments involve training on every dataset and evaluating on every other dataset for each task, resulting in 30 training runs per architecture. Multi-center experiments involve training on all datasets except one in leave=one-out fashion, also resulting in 30 runs, but with larger training sets. For disposition prediction experiments training is performed once for each model, as it is not a transfer study.

Overall, approximately $7 \cdot 12 \cdot (30 + 30 + 1) = 5124$ runs were performed. We use one to four top-of-the-line Nvidia H200 GPUs with up to 100GB of GPU memory for each run, depending on the task, architecture, and training set size. Multi-card training is done using `pytorch-lightning` [11]. Each compute server, an Nvidia Grace Hopper GH200 Superchip server, is equipped with up to 400GB of main memory, an ARM CPU with up to 288 cores, and has 4 GPUs. Experiments were run on a cluster infrastructure providing many servers with the aforementioned specifications. Each experiment shown in the paper was run on a single node (server).

E Impact and limitations

Impact. This work advances ML research for healthcare by enhancing models for early event prediction of adverse medical conditions. This research could in the future lead to improved care for patients in emergency units. This research incorporates data from multiple continents, making ML research in critical care time series more accessible and fair. Potential harmful impacts may include compromised patient safety. Investigation of the models to ensure their fairness, robustness, and privacy is an open topic for future works.

Limitations. This work has considered online early event prediction tasks as they are clinically relevant and typically harder than the alternatives. Other tasks can be considered (e.g., prediction of mortality, length of stay, sepsis) [34, 12]. We limit hyperparameter search to approximately 12 points per model due to huge computational burden of running the benchmark experiments (multiple seeds, multiple datasets, multiple tasks).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are included and supported with experiments where applicable.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See the relevant section.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [N/A]

Justification: This work does not present theoretical results.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [No]

   Justification: The paper uses well known architectures and describes the general structure of the pipeline. The code and harmonization tables are not yet fully released.

   Guidelines:

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

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

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

Answer: [No]

Justification: The paper does not release the code pending a broader study.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [No]

   Justification: The paper specifies most of the training and test details. The particular splits and hyperparameter sets are too numerous to include in the text.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [No]

   Justification: The experiments were made for multiple seeds, the paper does not present errorbars due to space constraints.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Information is provided on the scale of compute resources used and rough specifications of the servers without disclosing the specific architectures as it might disclose the computing center upon submission for double blind review.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: The research follow the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: See the relevant section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release data or models, so no safeguards are currently needed. The access to data is regulated and safeguarded by the providers.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper cites the sources of code and data used.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [N/A]

    Justification: This research does not involve human subjects (anonymized public access data).

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [N/A]

    Justification: This research does require IRB approval (anonymized public access data).

    Guidelines:

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