HiRID-ICU-Benchmark — A Comprehensive Machine Learning Benchmark on High-resolution ICU Data

Hugo Yèche [1] * **Rita Kuznetsova** [1] * **Marc Zimmermann** [1]

Matthias Hüser [1] **Xinrui Lyu** [1] **Martin Faltys** [1,2] **Gunnar Rätsch** [1]

{hyeche,mkuznetsova,marczim,mhueser,xlyu,mfaltys,raetsch}@inf.ethz.ch
[1]Department of Computer Science, ETH Zürich
[2]Department of Intensive Care Medicine, University Hospital, and University of Bern

Abstract

The recent success of machine learning methods applied to time series collected from Intensive Care Units (ICU) exposes the lack of standardized machine learning benchmarks for developing and comparing such methods. While raw datasets, such as MIMIC-IV or eICU, can be freely accessed on Physionet, the choice of tasks and pre-processing is often chosen ad-hoc for each publication, limiting comparability across publications. In this work, we aim to improve this situation by providing a benchmark covering a large spectrum of ICU-related tasks. Using the HiRID dataset, we define multiple clinically relevant tasks developed in collaboration with clinicians. In addition, we provide a reproducible end-to-end pipeline to construct both data and labels. Finally, we provide an in-depth analysis of current state-of-the-art sequence modeling methods, highlighting some limitations of deep learning approaches for this type of data. With this benchmark, we hope to give the research community the possibility of a fair comparison of their work.

Software Repository: https://github.com/ratschlab/HIRID-ICU-Benchmark/

1 Introduction

Severely ill patients require treatment and surveillance in Intensive Care Units (ICU). Critical health conditions are characterized by the presence or risk of developing life-threatening organ dysfunction. During a patient's stay in the ICU continuous monitoring of organs function parameters, enables early recognition of physiological deterioration and rapid commencement of appropriate interventions. Recent research shows the great success of machine learning methods when applied to ICU time series [44, 18]. One of the main goals of previous work was to develop new methods for prediction tasks relevant for clinical decision-making. Exemplary of such tasks are alarm systems that predict different types of organ failure [19, 41].

To develop and evaluate such methods only a small number of large-scale datasets are freely-accessible: The MIMIC-III [23] and IV [21] datasets, AmsterdamUMCdb [1], HiRID [11] and the eICU Collaborative Research Database [34]. However, these datasets are not provided in a pre-processed form directly suitable for machine learning nor do they have well-defined tasks, making it impossible to fairly compare works [22]. While some pre-processed alternatives with well-defined tasks exist [13, 35], they are often lacking in terms of size and diversity of tasks. We provide more

[*] Equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Track on Datasets and Benchmarks.

detail about this in section 2. This leads to situations where works compare methods on their private data [41] or only on limited data and number of tasks. Also the lack of relevant clinical sub-tasks for benchmarking stops the development of new methods for clinical decision support systems [15]. Finally, as in other fields, with recent datasets such as HiRID [11] the time resolution of data has greatly increased. However, no benchmark on ICU time series using such high-resolution datasets currently exists.

To improve this situation, in this paper we provide an in-depth benchmark based on the HiRID dataset [11, 19][2], which was released on Physionet [11] alongside the publication on the circulatory Early Warning Score (circEWS) [19]. HiRID is a freely accessible critical care dataset containing data recorded at the Department of Intensive Care Medicine, Bern University Hospital, Switzerland (ICU). The dataset was developed in cooperation with the Swiss Federal Institute of Technology (ETH), Zürich, Switzerland. We define a new benchmark on HiRID composed of various clinically relevant tasks and provide a comprehensive pipeline, which includes all steps from preprocessing to model evaluation. To assess the different aspects of the benchmarked machine learning methods, we diversify the tasks around specific challenges of ICU data such as prediction frequency, class imbalance, or organ dependency of the task. To profit from data acquisition advances and allow improvement on longer time series, we use a resampled data resolution of 5 min. HiRID has a higher time resolution than any other published critical care dataset and it motivates us to provide a comprehensive benchmark suite on this data-set. Also, we believe that this dataset will motivate the construction of new predictive methods for the healthcare field, going beyond ICU time series.

The main contributions of this paper are:

- We developed a comprehensive, end-to-end pipeline for time-series analysis of critical care data based on the recently published HiRID dataset. This pipeline includes the following stages: data preprocessing mode, training mode, and evaluation mode.
- We proposed and implemented a variety of tasks relevant to healthcare workers in the ICU, diversified in terms of type, prediction resolution, and label prevalence. The tasks cover all major organ systems as well as the general patient state. We included regression and classification (binary and multi-class) tasks.
- By providing a comprehensive benchmark on a set of canonical tasks, we give the research community around predictive modeling on ICU time series the possibility for the clear comparison of their methods.

The paper is organized as follows: in Section 2 we provide an overview of existing ICU datasets and benchmarking papers. We provide details about the HiRID dataset and introduce the tasks defined in collaboration with clinicians in Section 3 and give more details on the tasks in APPENDIX A: DATASET DETAILS. Section 4 describes the pipeline design, with more details given in APPENDIX B: HIRID-ICU-PIPELINE DETAILS. Section 5 describes the experiment and ablation study. In Section 6 we discuss the observed results and relate this paper to other benchmarks and related tasks relevant for clinicians.

2 Related Work

The main goal of this work is to provide a benchmark on the HiRID dataset for various clinical prediction tasks of interest. We describe here other ICU datasets as well as existing benchmarks for ICU data.

ICU time-series datasets There are several widely-used, freely-accessible datasets consisting of ICU time series. MIMIC-IV [23] is the oldest and most widely used ICU dataset available. It consists of physiological measurements as well as information about laboratory analyses. Physiological measurements are recorded with a maximum resolution of 1 hour. The results of laboratory analysis are collected at irregular time intervals. Moreover, there are static features like gender, age, diagnosis, etc. available. The dataset consists of information recorded about 40,000 ICU stays at Beth Israel Deaconess Medical Center (BIDMC), Boston, MA, USA. The median of the patient stay length is 2 days. The eICU Collaborative Research Database [34] is a large multicenter critical care database

[2] https://physionet.org/content/hirid/1.1.1/

made available by Philips Healthcare in partnership with the MIT Laboratory for Computational Physiology. It contains data associated with over 200,000 patient stays, but the public version does not reach the granularity of other datasets in terms of time resolution and data elements. The first version of AmsterdamUMCdb [1] was released in November 2019. Its current version from March 2020 contains data related to 23,172 ICU and high dependency unit admissions of adult patients from 2003 - 2016 from Amsterdam University Medical Centers. The data includes clinical observations like vital signs, clinical scores, device data, and lab results.

Benchmarks on ICU time-series. Among works using the openly available datasets mentioned above, to the best of our knowledge, only a single standardized benchmark exists, MIMIC-III benchmark by Harutyunyan et al. [15]. In that work four tasks were proposed, two requiring one prediction per patient stay and two dynamic tasks with one prediction per hour each. In addition, while they do not propose a benchmark, Jarrett et al. [20] developed a standardized pipeline for medical time series, called Clairvoyance. Results were shown on several datasets, including MIMIC. In this spirit, some packages address a specific family of tasks, for example, classification [12] and forecasting [14]. Finally, some public challenges, with curated data, were proposed in the past, e.g. for the early prediction of sepsis (Physionet 2019 challenge [35]) or mortality prediction (Physionet 2012 challenge [8]). However, the provided datasets are smaller than HiRID and built around a single task.

3 Benchmark Design

3.1 The HiRID Dataset

HiRID [11, 19] is a freely accessible critical care dataset containing data from more than 33,000 patient admissions to the Department of Intensive Care Medicine, Bern University Hospital, Switzerland (ICU) from January 2008 to June 2016. It was released on Physionet [11] alongside the publication of the circulatory Early Warning Score (circEWS) [19]. The dataset was developed in cooperation with the Swiss Federal Institute of Technology (ETH) Zürich, Switzerland. It contains de-identified demographic information and a total of 712 routinely collected physiological variables, diagnostic test results, and treatment parameters. HiRID has a higher time resolution than any other published ICU dataset, particularly for bedside monitoring, with most parameters recorded every 2 minutes, which motivates us to provide a comprehensive benchmark suite on this dataset. Demographic information about the patient cohort are displayed in Appendix Table 1.

Table 1: Definition of prediction tasks contained in the HiRID-ICU benchmark suite

Task name	**Task type**	**Task description**
ICU mortality	Binary classification, one prediction per stay	Predicted at 24h after admission to the ICU.
Patient phenotyping	Multi-class classification, one prediction per stay	Classifying the patient after 24h regarding the admission diagnosis, using the APACHE group II and IV labels[3]
Circulatory failure [4]	Binary classification, dynamic prediction throughout stay	Continuous prediction of onset of circulatory failure in the next 12h, given the patient is not in failure now.
Respiratory failure[5]	Binary classification, dynamic prediction throughout stay	Continuous prediction of onset of respiratory failure in the next 12h, given the patient is not in failure now.
Kidney function	Regression, dynamic prediction throughout stay	Continuous prediction of urine production in the next 2h as an average rate in ml/kg/h. The task is predicted at irregular intervals.
Remaining length of stay	Regression, dynamic prediction throughout stay	Continuous prediction of the remaining ICU stay duration.

3.2 Prediction Tasks

Our benchmark suite focuses on clinically relevant prediction tasks with a large diversity in the machine learning task types. The tasks cover most major organ systems as well as the general patient state. The major organ systems include the cardiovascular, kidney, and respiratory systems. For each organ system, we provide a prediction task related to the main organ function. Length of stay, mortality, and patient phenotyping are chosen to represent tasks regarding an overall patient state. From a machine learning point of view, our suite contains regression and classification (binary and multi-class) tasks. We included tasks with different degrees of class imbalance to diversify the spectrum further and enable the comparison of methods on e.g. highly imbalanced tasks. We chose tasks performed online throughout the stay (every 5 minutes) and at fixed times of the stay, such as 24h after ICU admission, which capture a more long-term state of the patient. To enhance reproducibility, we include some tasks previously considered in [15], mortality, and remaining length-of-stay prediction. Table 1 contains the full task descriptions.

4 Pipeline Design

Figure 1 shows an overview of the major HiRID-ICU pipeline steps. The pipeline is designed using the *preprocess-train-predict* paradigm. We provide more details about it in Appendix B: HiRID-ICU Pipeline Details and the README section of the software repository[6]. The preprocessed data contains two data versions, `common_stage` and `ml_stage`. The first is independent of modeling choices and serves as the starting point for future works with custom pre-processing choices. The latter is a compatible version for our pipeline with our categorical encoding, imputation, and scaling choices.

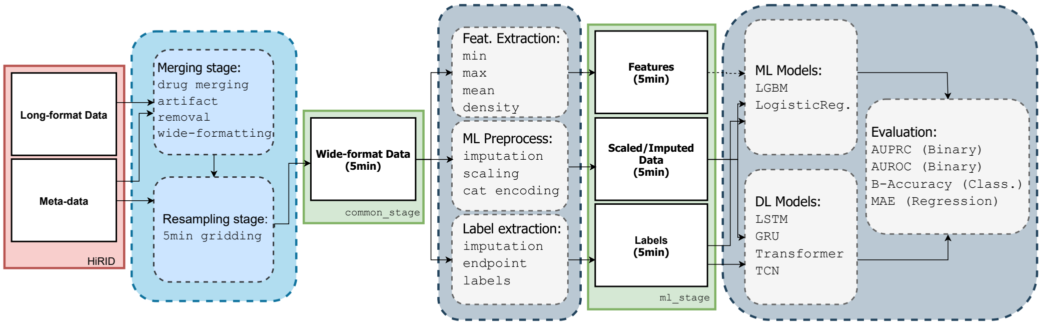

Figure 1: *Detailed Pipeline.* (Red) Raw Data. (Blue) Common pre-processing. (Green) Wide-format data. (Grey) Modeling depending stages.

4.1 Data Pre-processing

In its public version, HiRID, as any real-world dataset, contains certain artifacts that require pre-processing. As pointed out by [4] for MIMIC-III, individual pre-processing in each work avoids a fair comparison of them. To this effect, we aim to provide a modular and reproducible pipeline. Patient EHRs in HiRID are stored in a long table format where each row of the table is a record containing the measurement value of a specific variable at a specific time for a patient, which cannot be used as a ready input for machine learning models.

Wide-format Merging To obtain a more compact format, the first pre-processing step in our pipeline is to transform the long table of patient EHRs into feature matrices, where each column

[3] APACHE II and IV [46, 28] are subsequent versions of the major illness severity score used in the ICU. They also introduce a patient grouping according to admission reason. We use an aggregate of these two groupings for this task (see Appendix A: Dataset Details)

[4] Circulatory failure is defined as Lactate $>$ 2mmol/l and either mean arterial blood pressure $<$ 65mmHg or administration of any vasoactive drug.

[5] Respiratory failure is defined according to the Berlin definition [2] as a P/F ratio $<$ 300 mmHg.

[6] https://github.com/ratschlab/HIRID-ICU-Benchmark

represents a clinical concept, which we call wide format. Such a data format represents an irregularly sampled multivariate time series. At this step, we also remove any physiologically impossible measurement.

High-Resolution Gridding After this merging step, we further compact the dataset by re-sampling it to a 5 minute resolution. Thus, each time step contains the last value measured in the last 5 min if it exists, or is empty otherwise. This gridding is similar to the strategy used by [19]. We refer to the output of this step as the `common_stage` in Fig.1. Because it is independent of modeling choices, this stage provides a starting point for future approaches using different imputation and scaling choices.

Processing for Machine Learning In the second part of the pipeline, we process the common stage of the data to be compatible with ML models' expected input format. For this, we first use forward-filling imputation for each stay. Then, we apply one-hot encoding for categorical variables and scale the remaining ordinal or continuous variables. We standard scaled all variables with the exception of the time since admission and admission date, which we min-max scaled. By doing scaling globally, we ensure to preserve patients' specificity (e.g.: tachycardia). We refer to the output of this stage as the `ml_stage` as it is dependent on our modeling choices.

4.2 Hand-engineered Feature Extraction

In the original paper describing the HiRID dataset [19], the authors showed that boosted tree ensembles such as LGBM [25], when provided with hand-engineered features, outperform state-of-the-art deep learning methods. Based on this observation, we include in our pipeline the possibility to extract such features from the `common_stage` of the data. For our models, we extracted four features for each non-categorical variable over the entire history: *minimum past value*, *maximum past value*, *mean past value* and *density of measurement*, which is the proportion of time points where a value is provided among all possible time points in the history[7]. These features are then included in the `ml_stage`.

4.3 Label Construction and Splitting

We construct prediction task labels using the provided measurements and meta-data for both continuous and stay-level tasks. As an intermediate step for label construction, we use a forward imputed version of the data, as in the modeling stage. Concerning the experimental design, we use a random split by patients. The training set contains 70% of the patients and validation and test sets, 15% each. A temporal splitting strategy as used by Hyland et al. [19] would be more clinically relevant but information about admission time was removed to preserve anonymity when the dataset was originally published. While longer stays exist in the dataset, for computational reasons, we limited labeling to the first 7 days of stays (2016 steps of 5 min). This cropping affects less than 6% of all stays.

Table 2: Label statistics for each of the tasks, in the training, validation and test sets. As a metric, for binary classification tasks, the positive label prevalence is reported. For multi-class classification tasks, the class prevalence of the minority class is reported. Finally, for regression tasks the median of the label distribution is reported. In parentheses the number of samples is reported. M: Million.

Task name	**Train set**	**Validation set**	**Test set**
Circ. failure	1.4 % (n=14.12M)	1.3 % (n=3.01M)	1.4 % (n=2.96M)
Resp. failure	8.7 % (n=5.58M)	8.4 % (n=1.21M)	8.4 % (n=1.20M)
Mortality	8.7 % (n=10525)	7.1 % (n=2206)	8.3 % (n=2231)
Phenotyping	0.2 % (n=10470)	0.1 % (n=2194)	0.1 % (n=2217)
Kidney function	1.17 ml/kg/h (n=341424)	1.12 ml/kg/h (n=71549)	1.18 ml/kg/h (n=70642)
Rem. LOS	41.04h (n=15.15M)	41.51h (n=3.22M)	39.64h (n=3.17M)

[7] This is done on the regularly sampled version of the data

4.4 Model Training and Evaluation

The final part of the pipeline contains an end-to-end machine learning suite to train and evaluate our models depicted on the right hand side of Fig.1. ML approaches were implemented using `scikit-learn`[33] and `lightgbm`[25], whereas DL approaches were implemented in `pytorch` [32]. All DL models were trained using Adam optimizer [27], with a cross-entropy objective for classification tasks and mean-squared error (MSE) for regression tasks. For classification we provide the possibility to balance loss weights according to class prevalence as in [26].

For the evaluation of models, we use a range of metrics relevant to each task. For classification tasks, we considered AUROC[8] and AUPRC[9] metrics in the binary case, and balanced accuracy (B-Accuracy) [6] in the multi-class one. For regression tasks, we used mean absolute error (MAE) as a comparison metric. Regardless of the task or model, we used the `scikit-learn` implementation for all metrics. More details about this stage of the pipeline can be found in APPENDIX B: HIRID-ICU-PIPELINE DETAILS.

5 Experiments

5.1 Settings

For all models, we tuned specific hyper-parameters using random search. Each randomly picked set of parameters was run with 3 different random initializations. We then selected hyper-parameters on the validation set performance for either AUPRC, B-Accuracy, or MAE. All models were trained with early stopping on the validation loss. Further details about hyper-parameters can be found in APPENDIX B: HIRID-ICU-PIPELINE DETAILS.

Because of the class imbalance existing in classification tasks, we considered balanced loss weights for all methods. However as further discussed in subsection 5.5, this technique was relevant only for the Patient Phenotyping task. For regression tasks, we min-max scaled the labels at training time to avoid exploding gradients.

5.2 Benchmarked Methods

In our proposed benchmark, we considered two groups of machine learning algorithms. The first group consists of regular machine learning algorithms, which as shown are highly effective for ICU-related tasks [39, 15, 19]. It is composed of a Gradient Boosting method with LightGBM [25] and Logistic Regression. The second group is focused on deep learning methods. We select the most commonly used sequence models for this group: Recurrent neural networks (LSTM [16] and GRU [7]), convolutional neural networks (CNN), in particular, temporal convolutional networks (TCN) [3] and Transformer models [42].

5.3 Benchmarking Models on High-resolution ICU Data

In this section, we compare the earlier described methods on all tasks. While DL approaches are provided with the entire history for all time points, ML methods use only the values of the current step as an input. Thus one would expect the latter models to perform significantly worse due to the lower amount of information provided.

Stay-Level Tasks When comparing methods on tasks requiring a single prediction after 24h (Table 3), we observe the superiority of LGBM with hand-extracted features. Transformers outperformed other DL methods but we observe a significant performance gap with the best ML method in B-Accuracy for Patient Phenotyping and AUPRC for ICU Mortality. Concerning GRU and LSTM, their performance is similar to TCN's for ICU Mortality. However, on the Patient Phenotyping task, they do not manage to outperform even logistic regression.

Online Failure Predictions For the continuous classification tasks, where the maximum sequence length extends from 288 steps to 2016, DL methods do not leverage the additional history information.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

Table 3: *Benchmark of methods for stay level tasks*.(Top rows) ML methods; (Bottom rows) DL methods. All scores are averaged over 10 runs with different random seeds such that the reported score is of the form $mean \pm std$. In bold are the methods within one standard deviation of the best one. Classification metrics were scaled to 100 for readability purposes.

Task	ICU Mortality		Patient Phenotyping
Metric	AUPRC (↑)	AUROC (↑)	B-Accuracy (↑)
LR	58.1 ± 0.0	89.0 ± 0.0	39.1 ± 0.0
LGBM	54.6 ± 0.8	88.8 ± 0.2	40.4 ± 0.8
LGBM w. Feat.	**62.6** ± 0.0	90.5 ± 0.0	**45.8** ± 2.0
GRU	60.3 ± 1.6	90.0 ± 0.4	39.2 ± 2.1
LSTM	60.0 ± 0.9	90.3 ± 0.2	39.5 ± 1.2
TCN	60.2 ± 1.1	89.7 ± 0.4	41.6 ± 2.3
Transformer	61.0 ± 0.8	**90.8** ± 0.2	42.7 ± 1.4

Table 4: *Benchmark of methods for online monitoring tasks*. (Top rows) ML methods; (Bottom rows) DL methods. All scores are averaged over 10 runs with different random seeds such that the reported score is of the form $mean \pm std$. In bold are the methods within one standard deviation of the best one. Classification metrics were scaled to 100 for readability purposes. MAE is in units ml/kg/h for Kidney Function and in hours for Remaining LOS.

Task	Circulatory failure		Respiratory failure		Kidney func.	Remaining LOS
Metric	AUPRC (↑)	AUROC (↑)	AUPRC (↑)	AUROC (↑)	MAE (↓)	MAE (↓)
LR	35.6 ± 0.0	96.9 ± 0.0	49.3 ± 0.0	91.3 ± 0.0	N.A	N.A
LGBM	**43.5** ± 0.3	**97.7** ± 0.0	54.2 ± 0.3	92.8 ± 0.0	**0.45** ± 0.00	56.9 ± 0.4
LGBM w. Feat.	**43.9** ± 0.6	**97.7** ± 0.0	**55.7** ± 0.1	**93.1** ± 0.0	**0.45** ± 0.00	57.0 ± 0.3
GRU	41.9 ± 0.3	97.5 ± 0.0	52.8 ± 0.5	92.6 ± 0.1	0.49 ± 0.02	60.6 ± 0.9
LSTM	36.4 ± 1.3	96.6 ± 0.2	51.6 ± 0.8	92.3 ± 0.1	0.50 ± 0.01	60.7 ± 1.6
TCN	41.5 ± 0.5	97.5 ± 0.0	53.4 ± 0.4	92.7 ± 0.1	0.50 ± 0.01	59.8 ± 2.8
Transformer	42.1 ± 0.6	97.6 ± 0.0	53.7 ± 0.4	92.7 ± 0.1	0.48 ± 0.02	59.5 ± 2.8

Indeed, as shown in Table 4, for both Circulatory and Respiratory Failure, LGBM trained only on the current variables outperforms all DL methods. Among these methods, LSTM is the most impacted, as it has noticeably lower scores. Finally, for all continuous tasks, including regression discussed below, the improvement brought by hand-extracted features is not as significant. It suggests that statistical features, when extracted from the entire history, are less informative.

Online Regression Tasks The final set of tasks we benchmark are regression tasks (Table 4). As for the classification case, LGBM-based methods outperform DL methods, which, among them, have similar performance. In addition, we do not observe any improvement brought by our selection of hand-extracted features. Moreover, the overall performances of the proposed methods are relatively low. While a MAE of 0.45 ml/kg/h for Kidney Function is only two times smaller than the median urine output rate, a 57h error in Remaining LOS is more than two times the median length-of-stay. We believe these low scores are due to the nature of the labels' distributions, which are both heavy-tailed as shown in APPENDIX A: DATASET DETAILS.

5.4 Behaviour of Deep Learning Approaches for Long Time-Series

One notable difference between the MIMIC-III benchmark [15] and our work is the data resolution. The resolution of our data being twelve-time higher leads to 2016 steps (1 week) sequences for online tasks. Thus, we explore if the increase of sequence length explains DL methods' decrease in performances for continuous tasks.

History Length One way to verify if DL methods leverage long-term dependencies in their prediction is to check if a decrease in the considered history impacts performance. Among the DL methods, we can verify this for Transformers and TCN by using local attention in the first case and reducing the number of dilated convolutions in the other. In the results (Figure 2), we observe that these two models do not use the additional information provided by early steps. With only a 1h history, Transformer performance stays the same. TCN performance, on the other hand, does drop, but only for a history smaller than 6h. It could explain why LGBM, which does not have access to these extra time steps, manages to outperform both methods compared to stay-level tasks.

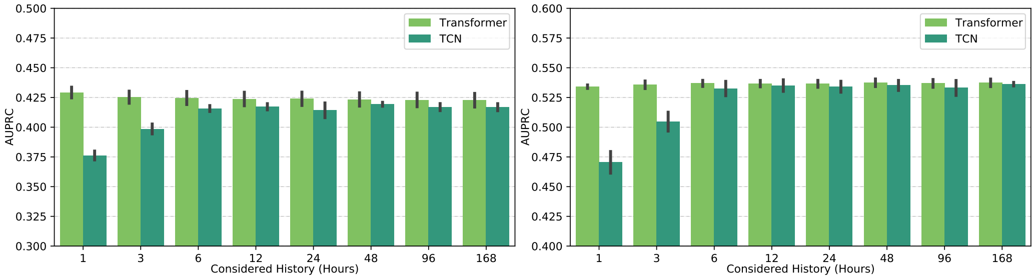

Figure 2: *Impact of history length on online classification performance.* (Left) Comparison in AUPRC for the Circulatory Failure task ; (Right) Comparison in AUPRC for the Respiratory Failure task. Error bars represent the standard deviation over 5 runs with different random initializations.

Data Resolution Another approach to decrease the length of sequences is to reduce the data resolution. We compare all DL methods with a 1h prediction interval to assess data resolution impact on metrics. This way, we can gradually decrease the data resolution from 5min to 1h while preserving the same prediction time-steps. We report the result of this experiment in Figure 3. We observe that while TCN and Transformer performance are almost identical, GRU and LSTM are both impacted in opposite ways. GRU is noticeably better than LSTM on both tasks with a 5min grid, but as resolution lowers to 1h, methods' performances become similar.

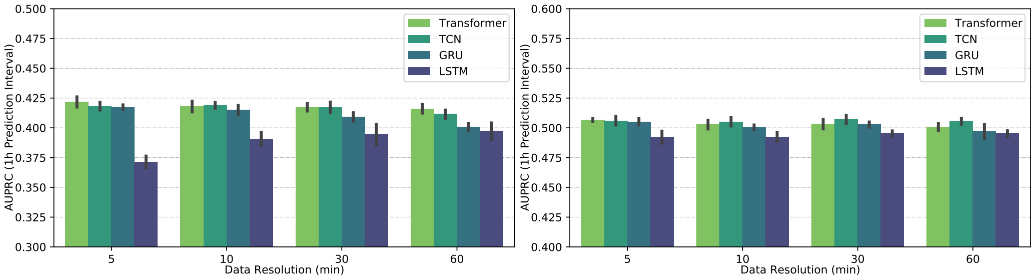

Figure 3: *Impact of data resolution on online classification performance.* (Left) Comparison in AUPRC for the Circulatory Failure task ; (Right) Comparison in AUPRC for the Respiratory Failure task. Error bars represent the standard deviation over 5 runs with different random initializations.

5.5 On Weighting Cross-entropy by Class Prevalence

All the tasks we define show a high degree of imbalance and the class imbalance problem (CIP) is known to be highly challenging [24]. The most common approach to this problem is the use of class weights in the loss objective. For this ablation, we adopt the original idea from [26] by defining weights inversely proportional to class prevalence. However, as shown in Table 5, such a weighting choice decreases performance in all binary classification tasks.

6 Discussion

In this paper, we provided an in-depth benchmark on the HiRID dataset and evaluated the behaviour of machine learning models on various clinically relevant tasks developed in collaboration with intensive

Table 5: *Deltas in metrics of using balanced cross-entropy loss.* (Green) Improvements over using no weights; (Red) Deterioration over using no weights.

Task	ICU Mortality		Phenotyping	Circulatory Failure		Respiratory Failure	
Δ Metric	AUPRC (↑)	AUROC (↑)	B-Accuracy (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)	AUROC (↑)
LGBM w. Feat.	-1.3	0.0	+4.3	-0.8	-0.2	-4.4	-1.1
Transformer	-2.6	0.0	+4.0	-0.6	0.0	-0.5	0.0

care clinicians. Our primary contribution is a full and reproducible preprocessing and machine learning pipeline and benchmark tasks on a public intensive care dataset, a necessary prerequisite for reproducible and comparable research in the future. We further evaluate current state-of-the-art machine and deep learning algorithms on these tasks establishing a baseline to compare future methods against. We consider this our second major contribution.

This work confirms previous results [19], that conventional machine learning models (i.e. boosted ensembles of decision trees) outperform current deep learning approaches on medical time series problems. Based on the experimental results we found that deep learning models do not lead to the same breakthrough performance increases as in other domains (such as NLP [9] or Computer Vision [10]). We believe the sparsity of the data and the imbalance of labels in both regression and classification tasks play an important role in this. For classification tasks, building a specific objective for highly imbalanced tasks such as Focal loss [31] might be a potential area of research. For regression, recent work has shown some promising leads for heavy-tailed regression tasks [43]. Moreover, HiRID introduces a novel high-resolution aspect in ICU data, that needs to be correctly taken into account. Thus, as for other sequence data, one possible explanation could be that when trained with extremely long sequences, models can not use the extracted features in the most effective way [45]. In the case of Transformers, to force the model to learn and extract useful patterns, various kinds of improvements could be made [40]. In particular, *learnable patterns* could be incorporated [37].

Our work goes beyond previous ICU time-series benchmarks (e.g. [15]) by using a more diverse set of tasks and a data set with a higher time resolution. As discussed earlier the set of clinical prediction tasks is diverse regarding the assessed organ systems, prevalence, and task type. An important limitation of our study is that HiRID is currently not the most frequently used and hence known ICU data set.

This work facilitates the future development of machine learning methods and standardized comparison of their performance on a diverse set of predefined tasks. It could contribute to solving today's problem of machine learning on medical time series not being comparable due to each work's unique datasets, preprocessing, and tasks definition. We hypothesize that methods developed and successfully evaluated on these tasks can also be successfully transferred to other specific medical time series problems.

This work also fills the gap between proposed machine learning approaches and their applications to ICU tasks. As a concrete example, COVID-19 is a big challenge for ICU patient monitoring. Important issues in this context are the uncertainty of the patient's prognosis as well as the prediction of the disease progression. COVID-19 is known to cause respiratory failure [30], one of the tasks studied in our benchmark, which is also the main cause for ICU admission and death [29, 38, 36, 17]. During the current COVID-19 pandemic, e.g. first attempts to construct a Respiratory Failure prediction model were already done [5], however, the data is available only for a limited audience, limiting reproducibility.

7 Conclusion

In this paper, we proposed an in-depth benchmark on time series collected from an Intensive Care Unit (ICU). In collaboration with clinicians, we defined several tasks relevant for healthcare covering different critical aspects of ICU patient monitoring. We provide a reproducible end-to-end pipeline to derive both data and labels, and a training setup to evaluate the final performance. We hope that this benchmark facilitates the construction and evaluation of machine learning methods for ICU data, and encourages reproducible research in this field.

References

[1] AmsterdamUMCdb. https://amsterdammedicaldatascience.nl/.

[2] ARDS Definition Task Force, V Marco Ranieri, Gordon D Rubenfeld, B Taylor Thompson, Niall D Ferguson, Ellen Caldwell, Eddy Fan, Luigi Camporota, and Arthur S Slutsky. Acute respiratory distress syndrome: the Berlin Definition. *JAMA*, 307(23):2526–2533, June 2012.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[4] David Bellamy, Leo Celi, and Andrew L Beam. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149*, 2020.

[5] Siavash Bolourani, Max Brenner, Ping Wang, Thomas McGinn, Jamie S Hirsch, Douglas Barnaby, Theodoros P Zanos, Northwell COVID-19 Research Consortium, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. *Journal of medical Internet research*, 23(2):e24246, 2021.

[6] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

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

[8] Luca Citi and Riccardo Barbieri. Physionet 2012 challenge: Predicting mortality of icu patients using a cascaded SVM-GLM paradigm. In *2012 Computing in Cardiology*, pages 257–260. IEEE, 2012.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] 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.

[12] Johann Faouzi and Hicham Janati. pyts: A python package for time series classification. *Journal of Machine Learning Research*, 21(46):1–6, 2020.

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

[14] Ahmed Guecioueur. pysf: Supervised forecasting of sequential data in python, 2018. https://github.com/alan-turing-institute/pysf.

[15] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Jan C Holter, Soeren E Pischke, Eline de Boer, Andreas Lind, Synne Jenum, Aleksander R Holten, Kristian Tonby, Andreas Barratt-Due, Marina Sokolova, Camilla Schjalm, et al. Systemic complement activation is associated with respiratory failure in COVID-19 hospitalized patients. *Proceedings of the National Academy of Sciences*, 117(40):25018–25025, 2020.

[18] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR, 2020.

[19] Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3):364–373, 2020.

[20] Daniel Jarrett, Jinsung Yoon, Ioana Bica, Zhaozhi Qian, Ari Ercole, and Mihaela van der Schaar. Clairvoyance: A pipeline toolkit for medical time series. In *International Conference on Learning Representations*, 2020.

[21] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, 2020.

[22] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376. PMLR, 2017.

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

[24] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[25] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[26] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] W A Knaus, E A Draper, D P Wagner, and J E Zimmerman. APACHE II: a severity of disease classification system. *Crit. Care Med.*, 13(10):818–829, October 1985.

[29] Xu Li and Xiaochun Ma. Acute respiratory failure in COVID-19: is it "typical" ards? *Critical Care*, 24:1–5, 2020.

[30] Yan-Chao Li, Wan-Zhu Bai, and Tsutomu Hashikawa. The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. *Journal of medical virology*, 92(6):552–555, 2020.

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

[35] Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

[36] Safiya Richardson, Jamie S Hirsch, Mangala Narasimhan, James M Crawford, Thomas McGinn, Karina W Davidson, Douglas P Barnaby, Lance B Becker, John D Chelico, Stuart L Cohen, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *Jama*, 323(20):2052–2059, 2020.

[37] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021.

[38] Qiurong Ruan, Kun Yang, Wenxia Wang, Lingyu Jiang, and Jianxin Song. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from wuhan, china. *Intensive care medicine*, 46(5):846–848, 2020.

[39] Reza Sadeghi, Tanvi Banerjee, and William Romine. Early hospital mortality prediction using vital signals. *Smart Health*, 9:265–274, 2018.

[40] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*, 2020.

[41] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[43] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. *arXiv preprint arXiv:2102.09554*, 2021.

[44] Hugo Yèche, Gideon Dredsner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*. PMLR, 2021.

[45] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big Bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*, 2020.

[46] Jack E Zimmerman, Andrew A Kramer, Douglas S McNair, and Fern M Malila. Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit. Care Med.*, 34(5):1297–1310, May 2006.