# Predicting Circulatory System Deterioration in Intensive Care Unit Patients

Stephanie L. Hyland<sup>1</sup>, Martin Faltys<sup>2,\*</sup>, Matthias Hüser<sup>1,\*</sup>, Xinrui Lyu<sup>1,\*</sup>, Cristóbal Esteban<sup>1</sup>, Tobias Merz<sup>2</sup>, and Gunnar Rätsch<sup>1</sup>

> ETH Zurich, Switzerland firstname.lastname@inf.ethz.ch
> Bern University Hospital, Switzerland tobias.merz@insel.ch
> \* contributed equally, alphabetical order

Abstract. The deterioration of organ function in ICU patients requires swift response to prevent further damage to vital systems. Focusing on the circulatory system, we build a model to predict if a patient's state will deteriorate in the near future. We identify circulatory system dysfunction using the combination of excess lactic acid in the blood and low mean arterial blood pressure or the presence of vasoactive drugs. Using an observational cohort of 45,000 patients from a Swiss ICU, we extract and process patient time series and identify periods of circulatory system dysfunction to develop an early warning system. We train a gradient boosting model to perform binary classification every five minutes on whether the patient will deteriorate during an increasingly large window into the future, up to the duration of a shift (8 hours). The model achieves an AUROC between 0.952 and 0.919 across the prediction windows, and an AUPRC between 0.223 and 0.384 for events with positive prevalence between 0.014 and 0.042. We also show preliminary results from a recurrent neural network. These results show that contemporary machine learning approaches combined with careful preprocessing of raw data collected during routine care yield clinically useful predictions in near real time.

Keywords: intensive care- circulatory system- machine learning

# 1 Introduction

Despite the high level of monitoring in the ICU, it is often infeasible for doctors to continually monitor the state of all patients. Unanticipated deteriorations can be life-threatening and require swift response. Identifying imminent or likely deterioration in a timely fashion is therefore an important question[2], and is the objective of research into early warning systems. Such systems have historically been based on a small number of physiological variables[21], allowing for easy assessment at the bedside but potentially missing complex patterns preceding deterioration[14]. As hospitals proceed to digitise data collection and visualisation, there is an opportunity for predictive algorithms to operate in real-time on this data, providing decision support to caregivers. Hyland et al.

In this work, we describe a data-driven predictive model for circulatory system failure. We integrate continuous measurements from hundreds of physiological variables and treatment parameters, drawn from a dataset of 44,655 patients over 8 years comprising 553.18 years of patient data. Our system identifies patterns indicative of pending haemodynamic instability, using many more variables than a typical ICU physician could assess. Currently relying on an observational dataset for internal validation, once finalised, this system will be deployed in a Swiss ICU to clinically validate its use as a real-time monitoring system.

#### 1.1 **Related Work**

Risk stratification on the basis of physiological parameters is a common practice in the ICU, and scores such as SOFA[22] explicitly quantify circulatory system dysfunction. SOFA and other scores (e.g. APACHE[12]) primarily draw on data from the first 24 hours in the ICU with an emphasis on mortality prediction. although repeated evaluation of SOFA has also been studied [6]. Mortality prediction has attracted interest from the machine learning community, producing benchmarking tasks [18,9] and modelling approaches such as ensembles [17], deep learning[3], and topic modelling[7]. In our case, the focus on real-time deterioration prediction puts the work closer in spirit to that of earning warning scores (e.g. MEWS[21]), which attempt to identify patients at risk of, for example, unplanned admission to ICU. Machine learning has also been exploited for the problem of ICU admission prediction, as in [1], and other early warning systems<sup>[4]</sup>. In this work, deterioration refers to the decline in function of the circulatory system in patients who are already in the ICU. [5] predict hyperlactatemia in ICU (MIMIC) patients, [20] predict hypotension using hidden Markov models, while [8] predict the onset of vasopressor usage.

#### 2 Data preparation

Preparing the data for use in a machine learning system was a critical component of this work. Raw data was exported from the patient database management system deployed at Bern University Hospital and then processed in several steps. Routinely-collected data of this kind features many challenges for computational analysis. Errors in data labelling (for example venous versus arterial blood gases), missing and implausible values, ambiguous or contradictory records, as well as artefacts introduced by routine care (for example blood pressure spikes due to arterial line flushing) necessitate care during data processing. In this work we attempted to remove the suspected erroneous data using variable-specific processing, resolving or deleting ambiguous records, and removing values based on plausible physiological ranges.

To deal with missing data, we compute the median sampling interval for each variable from training data, and forward-fill up to this point. After this, we decay to a rolling local median from the recent past (calculated over a similar interval). This reflects the belief that frequently-measured variables vary rapidly and should not be forward-filled for long, while decaying to the recent median

2

value implies that in the absence of data, we assume the patient has returned to 'baseline' (for them), where this can vary throughout their stay.

Medications were converted from doses to flow rates, treating 'instantaneous' drugs (such as tables) as flows over an effective active period, which we defined for each drug. In the database system used by the ICU at Bern University Hospital, drugs often received multiple unique identifiers for different dosage options, corresponding to different variables in the database. To address this and other redundancies in the data (for example, three ways of measuring temperature) we performed a manual dimensionality reduction step, merging variables that we identified to be sufficiently similar. In doing so, we reduced the total number of variables from 728 to 209. This means the model is applicable in any system measuring these variables, and is not specific to the ICU in Bern.

We use pandas[13], numpy[15], and scikit-learn[16] in Python for data processing and model development.

### **3** Deterioration prediction

We define deterioration as the appearance of a 'worse' state during a window up to  $\Delta t$  hours in the future. A patient can be in one of four states, where 0 is the best (stable), and states 1-3 describe increasing levels of circulatory system dysfunction. This dysfunction is identified through impaired circulatory function and elevated lactate values ( $\geq 2 \text{ mmol/L}$ ). Impaired circulatory function requires either low ( $\leq 65 \text{ mmHg}$ ) mean arterial pressure (MAP) or the presence of vasoactive drugs. To minimize spurious calls[19], we require these conditions to be true for at least 30 non-consecutive minutes of a 45-minute window.

The three dysfunctional levels are defined by the type and intensity of vasoactive drugs:

level	drugs requirement
1	Any dose of dobutamine, milrinone, levosimendan, or theophylline
2	$< 0.1 \mu g/kg/minute$ of norepinephrine or epinephrine
3	$\geq 0.1 \mu g/kg/minute$ of norepinephrine or epinephrine, or any dose of vaso pressin

The task is then binary classification on whether a patient in state s at time t will be in a state  $s + \delta_s$  ( $\delta_s > 0$ ) during a window starting five minutes from t and ending  $t + \Delta t$  hours later. We consider  $\Delta t$  in increments of one hour up to eight hours, the duration of a shift. For  $\Delta t = 8$ , this means the model can flag patients who may need additional attention during the next shift, but who may not be imminently critical.

As a model, we use an ensemble approach of boosted decision trees in the LightGBM library[11], with 200 trees and default hyperparameters otherwise. As this model does not natively handle time-series data, we generate derived features using five-point summary statistics (reporting min, max, median, interquartile range, and trend) over four temporal resolutions. The temporal resolutions depend on the sampling interval of the variable, allowing us to capture data from up to 72 hours in the past for slowly-varying parameters, and up to 12 hours in the past for higher-frequency variables. Other features include time

4 Hyland et al.

since admission, and fraction of time spent in circulatory failure so far. We also include early results from an LSTM[10] with hidden size 268, provided with at most *four hours* of (un-summarised) data.

In our experimental setup, we use the most recent six months of data to construct the test set, reflecting that such models are necessarily trained on retrospective data, and will be applied on new patients with a slightly different data distribution. We report AUROC as well as area under the precision-recall curve as deteriorations are relatively rare (1.4% prevalence during a one-hour window). The results for varying  $\Delta t$  are shown in Figure 3. We see that AUROC remains high as  $\Delta t$  increases, indicating high accuracy for predictions over the next shift. AUPRC is more challenging for this model and task, although performance is well above baseline (dotted line) for all  $\Delta t$  (between 9.14x and 15.93x). Given the preliminary nature of the LSTM results and the limited input data it receives (at most four hours), its performance is promising.



Fig. 1. AUROC as a function of deterioration horizon. Deterioration can occur during the window starting five minutes from now and ending at  $\Delta t$  hours.

Fig. 2. AUPRC as a function of deterioration horizon. Dotted line shows the prevalence of positive labels (deteriorations), which increases as the window size increases.

### 4 Conclusion

We have shown how careful handling of a large retrospective cohort of ICU patients results in a predictive model of circulatory organ failure with high (AU-ROC > 0.9) performance on time-horizons up to 8 hours in the future. We are currently developing models based on recurrent neural networks to better exploit the temporal nature of this data, and studying the behaviour of our trained classifier to identify ways to enhance positive predictive value. One direction is to provide the recurrent neural networks with longer history of data, such as the entire patient stays, so that the models can use as much information as available to make more accurate predictions. Once satisfactory *in silico*, this model will be deployed in the ICU for external validation. This work demonstrates the potential for large-scale multivariate modelling to identify patterns in physiological signals, enabling early warning of circulatory system deterioration.

## References

- Alaa, A.M., Yoon, J., Hu, S., van der Schaar, M.: Personalized risk scoring for critical care patients using mixtures of gaussian process experts. CoRR abs/1605.00959 (2016)
- Bates, D.W., Zimlichman, E.: Finding patients before they crash: the next major opportunity to improve patient safety. BMJ quality & safety 24 1, 1–3 (2015)
- Che, Z., Purushotham, S., Khemani, R., Liu, Y.: Interpretable deep models for icu outcome prediction. In: AMIA Annual Symposium Proceedings. vol. 2016, p. 371. American Medical Informatics Association (2016)
- Clifton, L.A., Clifton, D.A., Pimentel, M.A.F., Watkinson, P.J., Tarassenko, L.: Gaussian process regression in vital-sign early warning systems. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society pp. 6161–6164 (2012)
- Dunitz, M., Verghese, G., Heldt, T.: Predicting hyperlactatemia in the mimic ii database. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. pp. 985–988. IEEE (2015)
- Ferreira, F., Bota, D.P., Bross, A., Mélot, C., Vincent, J.: Serial evaluation of the sofa score to predict outcome in critically ill patients. JAMA 286 14, 1754–8 (2001)
- Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., Szolovits, P.: Unfolding physiological state: Mortality modelling in intensive care units. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 75–84. ACM (2014)
- 8. Ghassemi, M., Wu, M., Hughes, M.C., Szolovits, P., Doshi-Velez, F.: Predicting intervention onset in the icu with switching state space models. In: CRI (2017)
- Harutyunyan, H., Khachatrian, H., Kale, D.C., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:1703.07771 (2017)
- Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. pp. 3149–3157 (2017)
- Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E.: Apache ii: a severity of disease classification system. Critical care medicine 13(10), 818–829 (1985)
- 13. McKinney, W., et al.: Data structures for statistical computing in python
- Moss, T.J., Lake, D.E., Calland, J.F., Enfield, K.B., Delos, J.B., Fairchild, K.D., Moorman, J.R.: Signatures of subacute potentially catastrophic illness in the icu: Model development and validation. Critical care medicine 44(9), 1639–1648 (2016)
- 15. Oliphant, T.E.: A guide to NumPy, vol. 1 (2006)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research 12(Oct), 2825–2830 (2011)
- Pirracchio, R., Petersen, M.L., Carone, M., Rigon, M.R., Chevret, S., van der Laan, M.J.: Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. The Lancet Respiratory Medicine 3(1), 42–52 (2015)
- Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmark of deep learning models on large healthcare mimic datasets. arXiv preprint arXiv:1710.08531 (2017)

- 6 Hyland et al.
- 19. Schmid, F., Goepfert, M.S., Reuter, D.A.: Patient monitoring alarms in the icu and in the operating room. Critical care **17**(2), 216 (2013)
- Singh, A., Tamminedi, T., Yosiphon, G., Ganguli, A., Yadegar, J.: Hidden markov models for modeling blood pressure data to predict acute hypotension. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. pp. 550–553. IEEE (2010)
- 21. Subbe, C., Kruger, M., Rutherford, P., Gemmel, L.: Validation of a modified early warning score in medical admissions. Qjm **94**(10), 521–526 (2001)
- 22. Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., Thijs, L.: The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure (1996)