The Role of Context in the Prediction of Acute Hypotension in Critical Care

Niranjani Prasad
Princeton University
np6@princeton.edu

Konstantina Palla
Microsoft Research, Cambridge UK
kopalla@microsoft.com

Abstract

Applying machine learning tools to forecasting adverse events in intensive care can be invaluable in providing clinicians with the time needed to intervene and improve patient outcomes. In this work, we describe an end-to-end approach to the prediction of hypotension from critical care data using off-the-shelf classification models. Standard performance metrics suggest these models effectively learn from available data, and that additional multi-modal information improves classification accuracy. However, we show that this improvement is disputable when probing further into medical context and choices in data curation, thus highlighting the need for a domain-centric design of machine learning for clinical decision support.

1 Introduction

The democratization of machine learning practice over recent years, allowing application in a range of real world settings, raises both opportunities and unprecedented challenges [12]. The domain of healthcare epitomizes this; machine learning tools have the capacity to leverage the wealth of data available in electronic health records (EHRs) to support clinical decision making and improve efficiency. At the same time, great care and continual input from domain experts at every stage of the data selection and algorithm design is crucial to minimizing the risks of biased or confounded inferences [5]. In this work, we consider how some of these issues can arise through the lens of forecasting adverse clinical events. We focus on the task of predicting the onset of hypotensive events given raw physiological time series as input, along with context in the form of curated EHR data. We cast this as a binary classification problem and consider: (i) the trade-offs inherent in cohort and feature selection from noisy, multi-resolution, multi-modal data, (ii) the effectiveness of off-the-shelf machine learning algorithms in adverse event prediction, and (iii) the challenges in evaluating machine learning in the safety-critical domain of clinical decision making.

The paper is structured as follows: in Section 2, we introduce the problem of predicting hypotensive events and review past approaches to this task. In Section 3, we outline our data selection and preparation process and formulate our classification problem, while in Sections 4, we analyse the performance of baseline models trained and tested on a publicly available critical care dataset.

2 Background: Hypotension in Intensive Care

Hypotension is defined as a period of sustained, abnormally low blood pressure. It can not only be a harmful condition in itself, slowing the delivery of oxygen and nutrients to vital organs, but is often the first marker of more serious illness [3]. Acute hypotensive events (AHEs) are highly common in critical care patients, though estimating prevalence is challenging due to variability in clinical definitions across care providers, and across patient subpopulations; patients that meet a given population-level definition may be entirely asymptomatic, for example. An AHE can result from a number of different mechanisms, from distributive shock, typically caused by sepsis (severe blood

infection) or neurogenic disorders, or hypovolaemic shock following sudden loss of fluid, to shock directly due to heart or circulatory failure [18]. Hypotension is in turn associated with higher rates of comorbidity and mortality [8, 20, 16]. Timely prediction of hypotensive episodes can therefore allow clinicians to intervene as appropriate before further patient decompensation, and improve outcomes.

Existing literature on the prediction of AHEs focuses on the extraction and analysis of waveform shape and spectral characteristics, and the construction of complex hand-engineered features from raw waveform data with high temporal resolution [17, 1, 9]. This heavily featurized waveform data is then input to simple classifiers, such as logistic regression, random forests or support vector machines. Feature engineering becomes prohibitive when the number of the features increases and new combinatorial features become progressively difficult to interpret. Contrary to that, we deploy and explore baseline approaches that take a small set of physiological signals as input and relegate the composition of features to the classifier thus allowing for interpretability of the contribution of each signal to the prediction. Additionally, while past approaches to AHE consider only waveform data, here we look to incorporate information from corresponding EHRs as input.

More generally, the task of predicting adverse events ahead of clinical diagnoses has been tackled in a number of different contexts, from tree-based methods in predicting hypoxaemia [15] or building early warning systems for circulatory failure [10], to the detection of sepsis or acute kidney injury using recurrent neural networks [2, 19]. These works typically consider only sparse, irregularly sampled EHR data over extended time intervals; in the case of acute hypotension however, where both deterioration and treatment can occur at much shorter time scales, leveraging high fidelity waveform data—alongside clinical records as appropriate—is imperative.

3 Data Selection and Preprocessing

We train and evaluate models for the prediction task described using the MIMIC III Critical Care database [11], in conjunction with continual bedside monitoring data in the corresponding waveform database [6]. Motivated by prior approaches to AHE prediction [17, 4], we extract as features the time series of the following six vital signs, sampled once per minute: mean, systolic and diastolic arterial blood pressure, heart rate, respiratory rate and SpO_2 (blood oxygen saturation). We filter from the database a total of 4,518 patients with waveform data available for all six vitals. We define AHE onset as the point at which 80% of mean arterial pressure (MAP) measurements in the following 30 minute window are below 65mmHg, and label each patient admission according to whether a hypotensive event is present. This yields two groups of patients: cohort H, comprising 2,729 admissions with one or more instances of AHE, and cohort C, a control group of 1,789 patients who experience no AHEs.

We augment these six waveform time series with temporally aligned data from the clinical database. We consider information on administration of six categories of drugs that may directly influence blood pressure: vasodilators, diuretics, and sedatives can cause drops in blood pressure, while vasopressors and fluids (crystalloids and colloids) are administered to manage hypotension.Additionally, we extract timestamps of nurse-verified chart and lab events, indicative of suspected change in patient state, along with static demographic features (age, weight, gender, ethnicity, first care unit, admission type) that may help characterize expected patient baseline MAP.

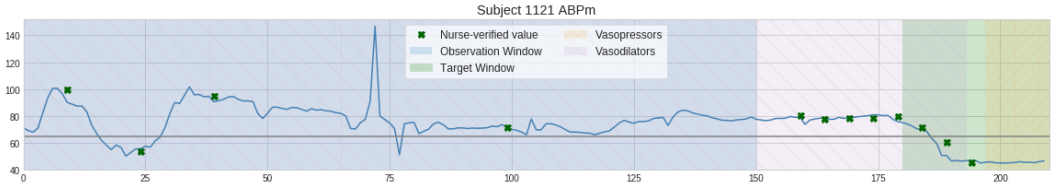

Figure 1: Example waveform time series segment of patient mean arterial blood pressure (ABPm), along with charted MAP values and administered drugs. AHE onset at $t^* = 180$; gap length $\Delta = 30$.

We extract fixed-length segments from each admission as follows: for patients in cohort H, we find the time t^* of the first AHE onset more than 180 minutes into the admission. We define the *observation window*, data in interval $[t^* - 180, t^* - \Delta]$. Here Δ is the *gap*, or the interval between the end of the observed data and the onset of the AHE event we aim to forecast, in the 30-minute *target window* $[t^*, t^* + 30]$. Figure 1 illustrates an example. For patients in cohort C, we simply extract the first $180 - \Delta$ minutes of data from each admission and use this as our observation window.

4 Experiments

Given the labelled dataset constructed above, we use information in the observation window of each sample to predict the probability of a hypotensive event occurring after time Δ, in the target window. We cast this as a binary classification task with class labels [H,C], to (i) evaluate the performance of common baseline classifiers, namely logistic regression, random forests and tree-based gradient boosting machine [13] when predicting AHE onset at different future intervals, taking the concatenated waveform time series of the six vitals as input, (ii) investigate the performance when explicitly modelling the temporal dynamics of the input using a stacked bidirectional LSTM classifier [7], and (iii) explore the effect of augmenting the classifier input with additional clinical information.

Figure 2(a) illustrates how the classification accuracy changes with interval lengths Δ, ranging from 0 and 60 minutes prior to AHE onset for each of our four classifiers. This can lend insight into the potential actionability of the predictions generated, taken in context of the response time of typical treatments for hypotension. As expected, we find that performance decreases with increasing Δ, though this typically plateaus after $\Delta = 30$, suggesting that a reasonable estimate of AHE risk can be achieved just with a short segment of data from the start of an admission.

We then consider how incorporating the auxiliary information from the clinical dataset can impact classification performance. Figure 2(b) suggests that inclusion the six time series of administered drugs consistently improves accuracy, while gains from chart measurement and lab test time series as well as inclusion of demographics are more modest, and in fact decrease accuracy in the case of logistic regression. This may be in part because demographics serve more as an indicator of baseline hypotension risk, rather than of immediate MAP deterioration. Figure 3 plots classifier ROC and precision-recall curve with $\Delta = 30$ and all extracted features, which can be used to choose a clinical operating point; for example, the GBM with false alarm threshold of 1 in 10 yields 91% recall.

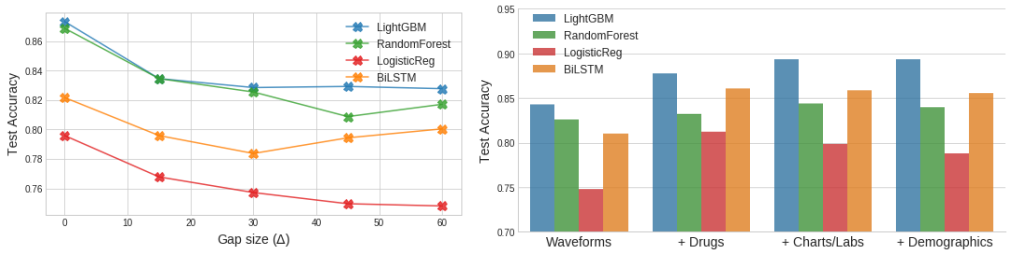

Figure 2: Comparison of classification accuracy of four baseline models, with (a) Sweep over different gap sizes Δ using waveform input alone; (b) Addition of clinical context features, fixed Δ = 30.

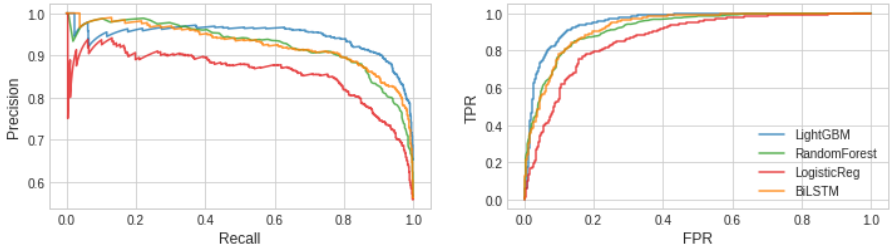

Figure 3: (a) Precision-Recall Curve, (b) Receiver-Operator Characteristic, for classification accuracy of each baseline model with varying thresholds, given all available features and gap size Δ = 30.

Interpreting feature importances In looking to explain the predictions of the GBM (the highest performing model), we use Shapley values for trees [14], which evaluate the contribution of each feature in pushing the predicted probability of AHE away from the population mean prediction, along with the direction of influence. In addition to population-level feature importances, it allows for individual-level explanations of predictions, important in facilitating trust in predictions. Figure 4 plots the distribution of the impacts each feature has on the model output, for the top 10 features, with color of sample point representing the feature value. We see that the top ten features are dominated by the value of diastolic (ABPd) or mean (ABPm) arterial pressure—where ABPm is a linear function of, and covaries with, ABPd—towards the end of the observation window. Samples with low values for these features tend to have high positive SHAP values, indicating increased probability of an AHE

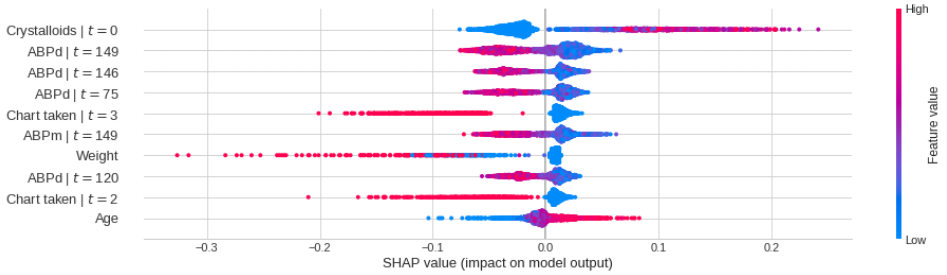

Figure 4: Top 10 features ranked by sum of *SHAP* value magnitudes over training samples, for GBM classifier with all extracted clinical features and gap $\Delta = 30$, such that | *observation window* | = 150.

in the target window, as would be expected. However, the feature with greatest impact across samples is crystalloids (fluids for blood volume expansion, increasing pressure) at the start of the observation window, $t = 0$. This suggests that the classifier is identifying those patients that have already been diagnosed by clinicians in the data as at high risk of hypotension, and hence have been administered fluids early in the admission. The classifier may therefore provide limited actionable insights in these cases, and predictions would be invalidated by any change in hypotension risk management practices.

This is emphasised when inspecting the highest impact features of instance-specific SHAP values for errors made by the classifier. False positives are dominated by admissions censored by preventative intravenous fluids; more crucially, false negatives often result from patients that do not receive fluids in the observation window and hence erroneously predicted as low-risk, despite deteriorating vitals. These issues motivate the need for causal approaches when building models with censored data.

Analysing patient subgroups We also explore how the distribution of errors varies with respect the patient comorbidities. Table 4 summarizes the performance of the GBM classifier across certain key patient subpopulations: patients admitted to cardiac surgery recovery or coronary care units (that are likely to be dependent on vasoactive drugs), patients that have been explicity diagnosed in ICD-9 codes with some form of shock, and those that expired in hospital. In each case, classification accuracy (recall in particular) are significantly higher for these subgroups that the whole test population. This suggests that our model performs better for more critically hyoptensive patients, and that many samples on cohort H may experience asymptomatic hypotension, of less clinical relevance.

Test Subgroup	# Admissions	$\#H \mid \#C$	Accuracy	Precision	Recall
Total	1318	729 | 589	0.894	0.934	0.88
Cardiac units	496	313 | 183	0.905	0.949	0.905
Non-Cardiac units	822	416 | 406	0.887	0.925	0.861
Shock (ICD-9)	241	183 | 58	0.913	0.951	0.935
No Shock (ICD-9)	1077	546 | 531	0.89	0.93	0.862
In-hospital mortality	179	135 | 45	0.905	0.963	0.915
Discharge	1139	595 | 544	0.892	0.929	0.872

Table 1: Sample size and classification accuracy for different patient subpopulations

5 Discussion

We presented an end-to-end approach to the prediction of hypotension from historical ICU data. We described our data selection process, and deployed a number of baseline classifiers to predict the onset of hypotensive events while varying input. We showed that, under standard performance metrics, off-the-shelf classification models perform well even in comparison with more sophisticated but data-hungry deep learning models, and this performance improves with the inclusion of information from multiple sources—that is, both raw physiological signals and data confounded by clinical action. We showed with further analysis, however, that accuracy provides a limited view of model usefulness, and evaluation in relation to clinical protocol underlying the collection of data, as well as decisions made in its curation, is crucial. This underlines the need for a principled, context-aware approach to the model design, with the synergy of medical and machine learning expertise.

References

[1] S. Bhattacharya, V. Rajan, and V. Huddar. A novel classification method for predicting acute hypotensive episodes in critical care. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 43–52. ACM, 2014.

[2] J. Futoma, S. Hariharan, and K. Heller. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1174–1182, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[3] M. Ghassemi. *Methods and models for acute hypotensive episode prediction*. PhD thesis, Oxford University, UK, 2011.

[4] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. pages 75–84, New York City, 2014.

[5] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, and R. Ranganath. Opportunities in machine learning for healthcare, June 2018.

[6] 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.

[7] A. Graves, S. Fernández, and J. Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*, pages 799–804. Springer, 2005.

[8] W.-J. Gu, B.-L. Hou, J. S. Kwong, X. Tian, Y. Qian, Y. Cui, J. Hao, J.-C. Li, Z.-L. Ma, and X.-P. Gu. Association between intraoperative hypotension and 30-day mortality, major adverse cardiac events, and acute kidney injury after non-cardiac surgery: A meta-analysis of cohort studies. *International journal of cardiology*, 258:68–73, 2018.

[9] F. Hatib, Z. Jian, S. Buddi, C. Lee, J. Settels, K. Sibert, J. Rinehart, and M. Cannesson. Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 129(4):663–674, 10 2018.

[10] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, et al. Machine learning for early prediction of circulatory failure in the intensive care unit. *arXiv preprint arXiv:1904.07990*, 2019.

[11] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[12] M. I. Jordan. Artificial intelligence—the revolution hasn't happened yet. 2019.

[13] 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. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[14] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019.

[15] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749, 2018.

[16] K. Maheshwari, B. H. Nathanson, S. H. Munson, V. Khangulov, M. Stevens, H. Badani, A. K. Khanna, and D. I. Sessler. The relationship between icu hypotension and in-hospital mortality and morbidity in septic patients. *Intensive care medicine*, 44(6):857–867, 2018.

[17] G. Moody and L. Lehman. Predicting acute hypotensive episodes: The 10th annual physionet/computers in cardiology challenge. volume 36, pages 541 – 544, 10 2009.

[18] A. Thompson. Hypotension: Issues and management. *The Pharmaceutical Journal*, 2011.

[19] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116, 2019.

[20] J.-L. Vincent, N. D. Nielsen, N. I. Shapiro, M. E. Gerbasi, A. Grossman, R. Doroff, F. Zeng, P. J. Young, and J. A. Russell. Mean arterial pressure and mortality in patients with distributive shock: a retrospective analysis of the mimic-iii database. *Annals of intensive care*, 8(1):107, 2018.