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

In this work, we explore how choices in data selection and algorithm design can confound predictions in machine learning for clinical decision-making. Focusing on hypotension prediction in intensive care, we consider the trade-offs inherent in cohort and feature selection from noisy, multi-resolution, multi-modal data, the effectiveness of off-the-shelf classifiers, and challenges in evaluation in this safety-critical domain.

Acute hypotensive events (AHEs)—sustained low mean arterial pressure—can occur due to sepsis or neurogenic disorders, shock from sudden fluid loss, or directly from cardiovascular disorders [10], and are in turn associated with higher mortality [2, 4, 11, 8]; timely prediction can therefore allow clinicians to intervene before further patient decompensation. Existing literature focuses on the construction of complex hand-engineered spectral features from high-resolution waveform data [9, 1, 5], which become progressively uninterpretable. Here, we use instead raw high-resolution waveform data from a small set of physiological signals, augmented by electronic health records. We train our models on 4,518 patients filtered from the MIMIC III Critical Care database [6], in conjunction with continual bedside monitoring data in the corresponding waveform database [3]. Patients were classified into two cohorts: those 2,729 admissions with one or more AHEs, and a control group. We extract a fixed-length segment from each admission (Figure 1): a target window for which we want to predict AHE risk, preceded by gap Δ, and an observation window—from which we extract waveforms of key vitals, temporally aligned data on fluids and vasoactive drugs, and demographic information—taken as our classifier input.

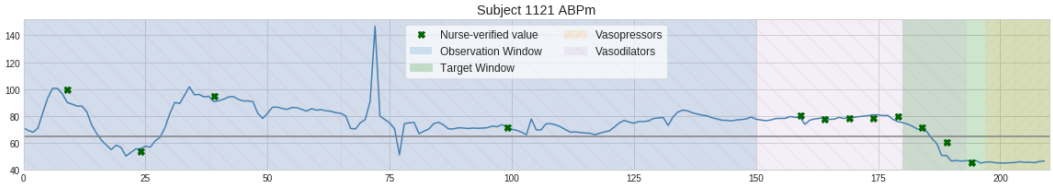

Figure 1: Example waveform time series segment of patient mean arterial blood pressure (ABPm), along with nurse-verified ABP measurement values and administered drugs. AHE onset (defined as a sustained 30-minute period of ABPm below 65mmHg) occurs at $t^* = 180$; gap length (i.e. prediction interval) $\Delta = 30$.

Figures 2 and 3 summarize how performance of several baseline classifiers varies with gap length and feature ablations. We use Shapley values [7] which evaluate the contribution of each feature towards the final prediction (enabling sample-level explanations) to analyse our best model. The top features (Figure 4) are dominated by the values of mean (ABPm) and diastolic (ABPd) arterial pressure towards the end of the observation window, as expected. However, the feature with greatest impact is crystalloid fluids (administered to expand blood volume and maintain blood pressure) at the start of the observation window. This suggests the classifier simply identifies those patients who are already diagnosed as at high risk of AHE—and administered fluids early—providing limited new, actionable insights.This is emphasised when inspecting classifier errors: false positives are dominated by censoring from preventative fluids; more crucially, false negatives often result from patients that do not receive fluids in the observation window and hence predicted as low-risk despite deteriorating vitals. Therefore, while off-the-shelf classification models perform well under standard performance metrics, these provide a limited view of model usefulness. Evaluation in relation to the clinical protocol underlying the collection of data, as well as decisions made in its curation, is crucial, and underlines the need for a principled, expert-in-loop approach to model design.

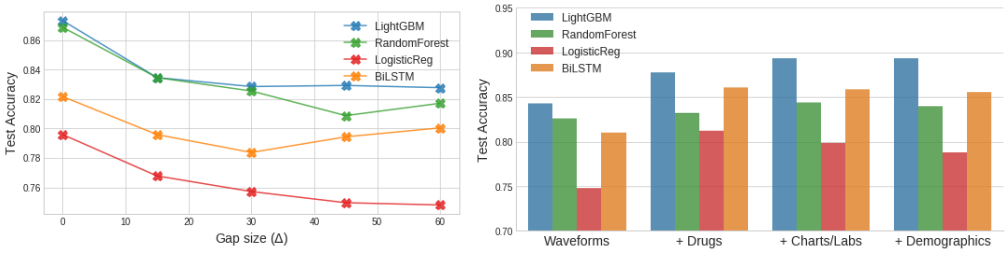

Figure 2: Comparison of classification accuracy of four baseline models: gradient boosted machines, random forests logistic regression, and bidirectional LSTMs, for (a) Sweep over different gap sizes Δ using waveform input alone; (b) Addition of clinical context features, fixed $\Delta = 30$. Performance decreases with increasing Δ, but 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. Inclusion of administered drugs consistently improves accuracy, while gains from chart measurement and lab test time series as well as demographics are more modest.

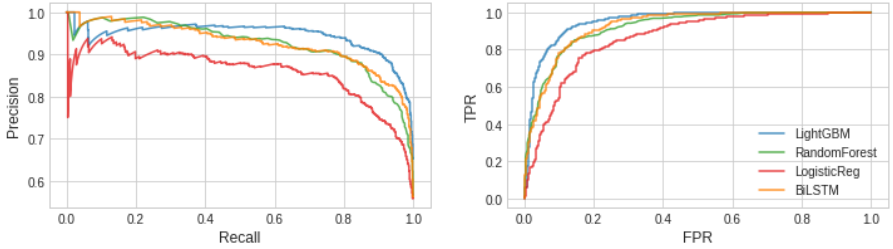

Figure 3: (a) Precision-Recall Curve, (b) Receiver-Operator Characteristic, for classification accuracy of each baseline model with varying thresholds, given all available features and gap size $\Delta = 30$.

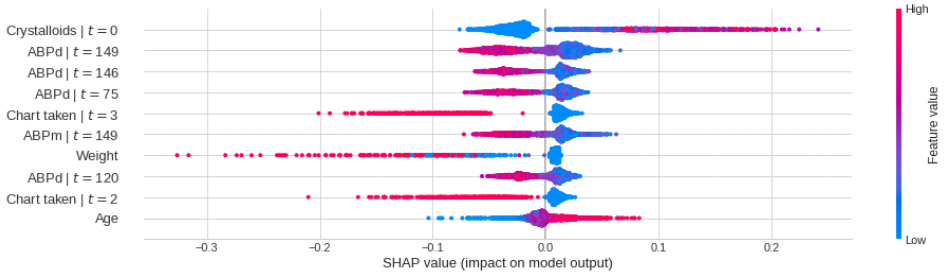

Figure 4: Top 10 features ranked by sum of *SHAP* value magnitudes over training samples, for GBM classifier with all extracted clinical features and gap $\Delta = 30$, such that $|$ *observation window* $| = 150$.. The color of each sample point represents the value of the feature in that sample.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.