Self-Supervised Early Warning of Critical Cardiorespiratory Events in Neonatal Vital-Sign Telemetry: A Pilot Feasibility Study

Craig Carlos Ouma*[1]

[1]Independent Researcher

June 9, 2026

Abstract

Continuous bedside monitoring generates dense vital-sign telemetry for hospitalised neonates, yet most of this signal is consumed by threshold alarms that fire only once an event is already underway. We investigate whether short-horizon *prediction* of critical cardiorespiratory events is feasible from routinely collected pulse-oximetry and temperature telemetry, using self-supervised labels derived from the signal itself rather than scarce clinical outcome labels. Using continuous recordings from three neonates (838,613 raw samples; 1,113 cumulative monitoring hours), we frame the task as predicting whether a desaturation event ($SpO_2 < 85\%$) will occur within a 60-second horizon, given a 30-second window of preceding telemetry. We evaluate a gradient-boosting baseline on engineered window features and a gated recurrent unit (GRU) on raw sequences, under strict leave-one-patient-out validation. The GRU achieves an area under the receiver operating characteristic curve (AUROC) of 0.70–0.96 across held-out patients, outperforming the gradient-boosting baseline (0.69–0.91). A leakage-safe lead-time analysis shows that flagged desaturations are warned a median of 80 s in advance, while exposing a gap between ranking performance and event-level sensitivity that ROC alone hides. We further show that a combined bradycardia-or-desaturation target does *not* generalise across this cohort, and trace the failure to the concentration of bradycardia in a single patient—a finding that directly informs cohort design for follow-up work. We position this as a proof-of-concept: with three patients the results establish feasibility and surface methodological constraints, but not population-level performance. Code and a reproducible training notebook are released.

1 Introduction

Neonates—particularly preterm infants—are monitored continuously for heart rate, respiratory rate, peripheral oxygen saturation (SpO_2) and temperature. Episodic desaturation and bradycardia are common and clinically meaningful: they reflect cardiorespiratory instability and, in aggregate, are associated with adverse outcomes [4, 3]. Conventional monitors act *reactively*, raising an alarm only once a vital sign has already crossed a threshold. A predictive system that anticipates such events even tens of seconds in advance could shift clinical response from reaction to pre-emption.

A natural but ultimately infeasible framing of this problem is direct mortality prediction. Mortality is a single, patient-level binary outcome; in a cohort of a handful of infants it provides at most

*Correspondence: craigcarlos95@gmail.com

a handful of labels, and dense per-second telemetry does not increase the effective sample size for a patient-level target. The dataset available here additionally carries no outcome field. We therefore deliberately avoid a mortality model and instead adopt a *self-supervised* formulation: the prediction targets are defined from future values of the same telemetry stream, yielding hundreds of thousands of labelled windows from a small number of patients.

This study asks a narrow, answerable question: *can a short window of neonatal vital-sign telemetry predict an imminent critical event, and does such a model generalise to an unseen patient?* We treat the work as a pilot feasibility study and are explicit throughout about what a three-patient cohort can and cannot support.

2 Background and Related Work

Two threads motivate the design. First, heart-rate-characteristic (HRC) monitoring established that subtle abnormalities in cardiac dynamics precede clinical deterioration in preterm infants, and a multi-centre randomised trial found that displaying an HRC index at the bedside reduced mortality in very low birth weight neonates [1, 2]; this is the clinical rationale for treating bradycardia as a high-value target. Second, a body of work on detecting and predicting apnea–bradycardia–desaturation events from neonatal vitals shows that the task is tractable with both tree-based and neural architectures, even under limited and class-imbalanced data [4, 3]. Our contribution is not a new architecture but an honest feasibility assessment on routinely collected, intermittently sampled device telemetry, with patient-aware validation and a transparent account of where small n limits the conclusions.

3 Data

The dataset comprises continuous "NeoMonitor" device recordings for three neonates (patient identifiers 18, 22 and 87), with no personal identifiers present. Each record contains a millisecond timestamp, pulse rate, respiratory rate, SpO_2, temperature, two signal-quality indices (`snrIr`, `snrRed`) and device alarm/error flags. Table 1 summarises the cohort.

Table 1: Cohort summary. Pulse-oximetry channels are present only when the optical sensor is actively reading; temperature is logged continuously.

	Patient 18	Patient 22	Patient 87
Raw samples	201,421	391,475	245,717
Monitoring span (h)	621.0	363.4	128.8
SpO_2 present (%)	16.8	19.2	13.0
Temperature present (%)	100	100	100
Mean pulse (bpm)	118	115	127

Three properties of the data shape the methodology. (i) **Sampling.** The nominal cadence is 1 Hz, but recordings contain long gaps where the device is detached; we therefore segment each record at gaps exceeding 10 s and never construct a window that crosses a gap. (ii) **Intermittent pulse oximetry.** Heart rate and SpO_2 are populated only while the optical sensor reads (13–19% of samples); we restrict the modelling set to samples where both are present. (iii) **Value persistence.** The device holds the last valid reading between updates—approximately 92% of consecutive SpO_2 values are identical—so absolute counts of "samples" overstate the number of

independent observations. The continuously logged temperature channel reads below 36.5°C for the large majority of the record, which is more consistent with probe placement and ambient conditions than with sustained core hypothermia; we include temperature as an input feature but do not interpret it clinically.

Event prevalence. Table 2 reports the prevalence of the two candidate events among valid pulse-oximetry samples. Desaturation occurs in every patient, whereas bradycardia is heavily concentrated in patient 22. This asymmetry is central to the results below.

Table 2: Per-patient event prevalence among valid pulse-oximetry samples.

Event	Patient 18	Patient 22	Patient 87
Bradycardia (HR < 100)	0.4%	5.7%	0.7%
Desaturation ($SpO_2 < 85$)	15.8%	6.8%	4.2%

4 Methods

4.1 Task formulation

Let t index valid pulse-oximetry samples within a contiguous segment. For an input window of $W = 30\,\mathrm{s}$ ending at t and a prediction horizon of $H = 60\,\mathrm{s}$, the label is

$$y_t = \mathbb{1}[\,\exists\,\tau \in (t, t+H] : \text{event occurs at } \tau\,],$$

where the event is desaturation ($SpO_2 < 85\%$), bradycardia (HR $< 100\,\mathrm{bpm}$), or their disjunction ("combined"). The headline analysis uses the desaturation target; the combined target is examined separately as a generalisation stress test.

4.2 Models

Gradient-boosting baseline. For each window we compute, per channel, the mean, standard deviation, minimum, maximum and a window-length slope, and fit a LightGBM [5] classifier with balanced class weights. This model is small, fast, and deployable on edge hardware.

Recurrent model. A two-layer gated recurrent unit (GRU; hidden size 64, dropout 0.2) [6] consumes the raw 30×6 window and is trained with a class-balanced binary cross-entropy objective. Per-channel normalisation statistics are computed from the training patients only, to prevent leakage across the validation boundary.

4.3 Validation

All evaluation is **leave-one-patient-out**: the model is trained on two patients and tested on the third, rotating across all three. This is the only patient-aware protocol available at $n = 3$ and is the appropriate test of whether a model transfers to an unseen infant. We report AUROC, area under the precision–recall curve (AUPRC), and the positive rate of the held-out fold, since AUPRC must be read against fold prevalence.

5 Results

5.1 Desaturation prediction

Table 3 reports leave-one-patient-out performance for the desaturation target. Both models predict imminent desaturation well above chance for two of three held-out patients, with the GRU outperforming the gradient-boosting baseline on every fold. Patient 87 is the hardest fold for both models; it has the lowest desaturation prevalence and the shortest record, leaving the least patient-specific signal for the two training patients to cover.

Table 3: Leave-one-patient-out performance for desaturation prediction ($SpO_2 < 85\%$ within 60 s). Positive rate is for the held-out fold.

Held-out patient	LightGBM		GRU	
	AUROC	AUPRC	AUROC	AUPRC
Patient 18	0.910	0.806	0.955	0.873
Patient 22	0.843	0.575	0.899	0.713
Patient 87	0.688	0.364	0.697	0.421

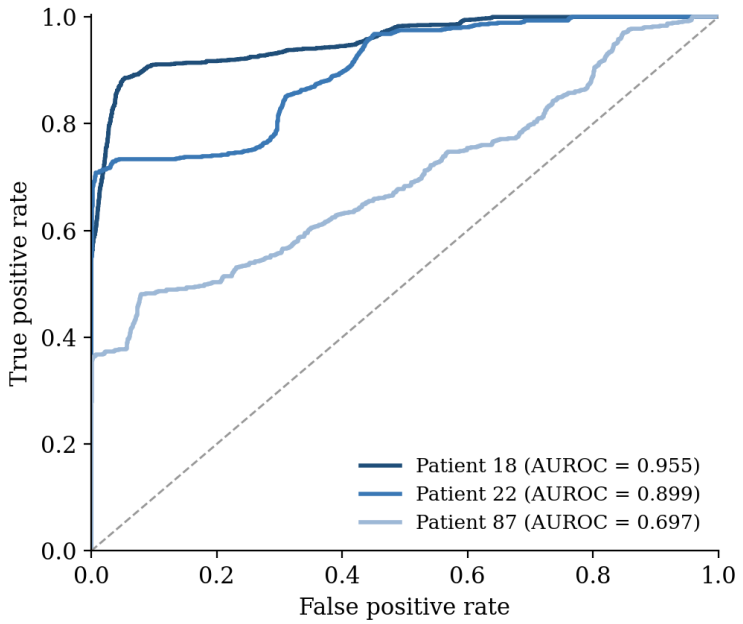

Figure 1: Leave-one-patient-out ROC curves for GRU desaturation prediction. Performance is strong for patients 18 and 22 and degrades for patient 87, the shortest record with the lowest event prevalence.

5.2 Lead-time analysis

AUROC measures ranking quality but not bedside readiness, which depends on how much warning a usable operating point provides. For each fold we fixed the alarm threshold on the *training* patients'

scores at 80% sensitivity (a leakage-safe choice) and applied it unchanged to the held-out patient. We then identified discrete desaturation onsets (transitions from $SpO_2 \geq 85$ to < 85) and, for each, measured the earliest alarm within the preceding 120 s.

Pooled across folds, 33 discrete onsets occurred; 13 were flagged in advance and 20 were not detected at this operating point. Among the flagged events the warning was substantial: median lead time 80 s (interquartile range 12–120 s), with 77% giving at least 10 s and 69% at least 30 s of warning (Figure 2). The mismatch between high AUROC and an event-level sensitivity of 13/33 (≈39%) is the central practical finding: ranking performance overstates clinical readiness, and at a conservative, calibration-free operating point a meaningful fraction of episodes are missed. The 120 s look-back also right-censors the lead-time distribution—several alarms were already active at the window boundary—so the median understates the achievable warning. Both observations point to threshold calibration on a held-out development set, and a larger cohort, as the necessary next steps.

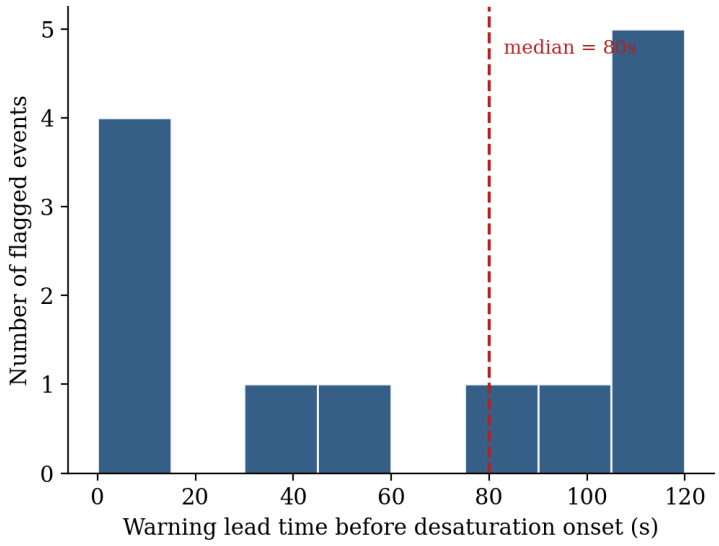

Figure 2: Distribution of warning lead times for the 13 desaturation onsets flagged in advance, pooled across leave-one-patient-out folds (120 s look-back). The mass near 120 s reflects right-censoring at the look-back boundary.

5.3 Combined target and the limits of $n = 3$

Replacing the desaturation target with the combined bradycardia-or-desaturation event degrades cross-patient generalisation sharply for the gradient-boosting model: the held-out AUROC for patient 22 falls to 0.489—indistinguishable from chance—while the other folds are largely unchanged. The explanation is in Table 2: bradycardia is almost entirely a patient-22 phenomenon. When patient 22 is the held-out fold, the model is trained on patients 18 and 87, which together contain very few bradycardia events, so the bradycardic component of patient 22's combined label is effectively unlearnable. This is not a modelling defect but a cohort-composition constraint: *cross-patient bradycardia prediction cannot be demonstrated until the cohort includes more than one bradycardic infant.* We report this explicitly because it is the single most important determinant of follow-up study design.

6 Discussion

The desaturation results establish feasibility: a short window of routine telemetry carries enough information to anticipate an imminent desaturation in an unseen neonate, and a modest recurrent model extracts it without hand-engineered features. The lead-time analysis sharpens this: when the model flags an episode it does so a clinically useful interval ahead (median 80 s), but at a conservative operating point it misses a substantial share of episodes. Bedside value therefore hinges as much on threshold calibration as on the underlying ranking model—a distinction that AUROC obscures and that any deployment claim must address.

For deployment, the gradient-boosting model is the pragmatic choice: it is compact, fast, and runs on the live stream without outcome labels at inference time. Crucially, an event-warning model can be deployed from data of exactly this form, whereas a mortality model cannot be built from it at all.

7 Limitations

This is a three-patient pilot. The results show feasibility and surface methodological constraints; they do not estimate population-level performance and should not be read as clinical validation. Specific limitations: (i) $n = 3$ precludes any population claim and makes each leave-one-out fold sensitive to a single patient's physiology; (ii) value persistence in the device stream inflates raw sample counts and induces autocorrelation that the patient-aware split controls for across patients but not within; (iii) the desaturation target is partly autocorrelated with recent SpO_2, so a portion of predictive performance reflects signal continuity rather than novel anticipation; (iv) bradycardia and combined targets are not validated cross-patient for the reasons above; (v) no external validation cohort is available. Addressing these requires a larger, more physiologically diverse cohort; on the order of 20–30 infants would begin to support population-level analysis and cross-patient bradycardia validation.

8 Conclusion

We have shown that short-horizon prediction of neonatal desaturation from routine, intermittently sampled telemetry is feasible and transfers to unseen patients in a three-patient pilot, with a GRU reaching AUROC up to 0.96 under leave-one-patient-out validation. We have also shown, transparently, where a three-patient cohort breaks down—bradycardia generalisation—and why. The work is offered as a reproducible proof-of-concept and a design specification for a properly powered follow-up study.

Reproducibility

The full preprocessing, feature construction, both models, and the leave-one-patient-out evaluation are released as a single notebook with a configurable event target (`desat`/`brady`/`combined`). All numbers reported here are produced by that pipeline.

Ethics and Data

The author confirms the right to publish these data. The recordings contain no personal identifiers; patients are referenced only by non-identifying study numbers.

References

[1] J. R. Moorman, W. A. Carlo, J. Kattwinkel, R. L. Schelonka, P. J. Porcelli, C. T. Navarrete, E. Bancalari, J. L. Aschner, M. W. Walker, J. A. Perez, C. Palmer, G. J. Stukenborg, D. E. Lake, and T. M. O'Shea, "Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial," *The Journal of Pediatrics*, vol. 159, no. 6, pp. 900–906.e1, 2011.

[2] K. D. Fairchild and T. M. O'Shea, "Heart rate characteristics: physiomarkers for detection of late-onset neonatal sepsis," *Clinics in Perinatology*, vol. 37, no. 3, pp. 581–598, 2010.

[3] B. A. Sullivan and K. D. Fairchild, "Predictive monitoring for sepsis and necrotizing enterocolitis to prevent shock," *Seminars in Fetal and Neonatal Medicine*, vol. 20, no. 4, pp. 255–261, 2015.

[4] A. Honoré, V. Siljehav, S. Chatterjee, and E. Herlenius, "Large neural network based detection of apnea, bradycardia and desaturation events," in *NeurIPS 2017 Workshop on Machine Learning for Health*, 2017. arXiv:1711.06484.

[5] 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 (NeurIPS)*, vol. 30, 2017.

[6] 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," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.