



**Machine Learning for Early Prediction of
Bradycardia in Preterm Infants:
A Lead Time Analysis on a Public
Cardio-Respiratory Dataset**

by

Mohammad Lakhi

Supervisor: Dr Nouh Elmitwally

Project submitted as part of the requirements for the
award of MSc Data and Artificial Intelligence

August 2026

Abstract

Preterm infants in neonatal intensive care are constantly monitored for bradycardia, but a significant percentage of the alarms their monitors produce are false. Therefore, a system that provided warning prior to an event rather than after would be more beneficial than one that merely reacts. Unlike most research in the field, this dissertation focuses on how far in advance bradycardia can be predicted rather than whether it can. It makes use of the Preterm Infant Cardio-Respiratory Signals database, which is the only publicly accessible dataset of its kind and includes 10 preterm infants, 439.8 hours of recording, and 622 expert annotated bradycardia onsets. An elastic net logistic regression and a one-dimensional convolutional network were trained under leave one patient out cross validation, and a windowing and feature extraction pipeline was constructed in accordance with the published design of Varisco et al. A direct measurement of the rate at which predictive information deteriorates prior to an onset was prompted by an initial protocol that used lead times of one, three, and five minutes and produced results that were indistinguishable from chance across three different model families. It deteriorates rapidly: The usable horizon for these two signals is seconds rather than minutes because discrimination drops from an AUROC of 0.92 at the moment of onset to 0.52 twenty seconds earlier. Since bradycardia is defined by heart rate, respiration timing provided more information than heart rate within that time frame, reaching an AUROC of 0.76 against 0.62 at a five-second lead. This result cannot be explained by circularity. In every lead time test, the linear model performed better than the neural network. The system’s sensitivity at the event level was 0.43, but its precision was only 0.12, far below the 0.35 baseline that the project aimed to surpass. Therefore, rather than a system prepared for clinical use, the contribution of this work is a measured account of what this reduced signal set can and cannot support.

TABLE OF CONTENTS

1	Introduction and Problem Statement	1
2	Related Work	4
2.1	Clinical context and prediction justification	4
2.2	The wider landscape: scope, taxonomy, and recurring limitations	4
2.3	Continuous physiological signal detection based methods	6
2.4	Prediction-based approaches: forecasting ahead of onset	7
2.5	Different demographics, modalities, and definitions of outcomes	8
2.6	Neonatal AI on a larger scale using a new data modality	9
2.7	Examined and disregarded studies	10
2.8	Research gap and synthesis	10
3	Methodology	13
4	Experimental Design	17
4.1	Rationale for the papers underpinning this design	17
4.2	Exploratory investigation preceding the fixed experimental protocol . .	18
4.3	Experimental variables	21
4.4	Hypotheses	22
4.5	Justification of key design decisions	22
4.6	Threats to validity	24
4.7	Tools and computational environment	25
5	Results	26
5.1	Window level performance by lead time and signal set	26
5.2	Comparison of the two models	27
5.3	Feature relevance	28

5.4	Performance at the event level	29
5.5	Inter-patient variation and event-level precision	31
5.6	Summary of hypothesis outcomes	33
6	Discussion	34
6.1	Principal findings	34
6.2	The lead time horizon	34
6.3	Respiration against heart rate	35
6.4	Comparison with previous work	36
6.5	Alarm burden and clinical implications	36
6.6	Limitations	37
6.7	Future work	38
6.8	Conclusion	38
	REFERENCES	40

Chapter 1

Introduction and Problem Statement

One of the most prevalent cardiorespiratory instabilities afflicting preterm newborns in neonatal intensive care is bradycardia, which often coexists with apnea and oxygen desaturation as part of the larger apnoea bradycardia hypoxaemia (ABH) pattern observed in this population. Although there is significant individual variability and no agreed upon minimum criteria for diagnosing apnea of prematurity in the first place, this immaturity of respiratory control is universal among infants born preterm, with severity inversely related to gestational age, and usually resolves as postmenstrual age advances [12].

Therefore, continuous physiological monitoring is essential to the practice of neonatal intensive care units (NICUs). However, a thorough scoping review of continuous cardiorespiratory monitoring in hospitalized children revealed that the vast majority of novel monitoring approaches and prediction models found in the literature are still small, single center, and infrequently validated prospectively or externally [1]. This issue is exacerbated by conventional threshold-based alarms, which are reactive rather than predictive, notify personnel only after a physiological threshold has been crossed, and are frequently linked to high false-alarm rates and the ensuing alarm fatigue [6, 7].

In an effort to create monitoring systems that are more accurate and truly predictive rather than merely reactive, a sizable and expanding body of research has applied machine learning (ML) to routinely collected cardiorespiratory signals, such as electrocardiography (ECG), respiration, and peripheral oxygen saturation [3, 4, 9, 15]. The majority of published approaches in this field still rely on traditional machine learning instead of deep learning, according to reviews of this field [3]. These reviews consis-

tently identify data quality, generalizability, interpretability, and limited prospective validation as the main unresolved challenges in the field [6, 19].

This dissertation examines whether an early-warning prediction model for bradycardia can be supported by cardiorespiratory signals from the Preterm Infant Cardiorespiratory Signals (PICS) database, a publicly accessible dataset of ECG and respiration recordings from ten preterm infants, annotated with bradycardia onset times. The annotation structure of PICS directly offers ground-truth onset dates, in contrast to a number of previous research that categorize candidate events previously reported by a monitor as true or false alarms [4, 9]. Instead of categorizing an event that has previously been identified, this provides a true forecasting task: estimating, from a previous window of information, if a bradycardia onset is likely within a specified lead time.

The specific objectives of this dissertation are as follows:

1. To critically evaluate the literature on machine-learning-based detection and prediction of bradycardia and associated apnoeic events in preterm infants, and to identify the specific research gap this work addresses.
2. To construct a feature-extraction and windowing pipeline suited to the annotation structure of the PICS dataset, following the 30-second moving-window scheme of Varisco et al. [9, 4].
3. To determine empirically how far in advance of a bradycardia onset any predictive information is present in this dataset, rather than assuming a lead time a priori.
4. To train and evaluate two machine learning models, elastic-net logistic regression and a one-dimensional convolutional neural network, under leave-one-patient-out cross-validation at the lead times identified in Objective 3.
5. To quantify how much predictive performance is retained when only respiration-derived features are available, given that Zuzarte et al. [15] report near-chance performance for cardio-respiratory features in the absence of movement-derived features.

6. To report performance honestly at both window level and event level, including false-alarm burden, and to situate it against previous work only where such comparison is scientifically valid.

Chapter 2

Related Work

2.1 Clinical context and prediction justification

Immature central regulation of breathing is the cause of apnea of prematurity and its accompanying bradycardia, a disorder that is common to premature delivery but varies in severity, resolution time, and long-term consequences [12]. Although infants born at the lowest gestational ages may continue to show shorter breathing pauses, intermittent hypoxia, and bradycardia much later, often delaying hospital discharge, clinical resolution usually occurs by 34-36 weeks postmenstrual age [12]. Unresolved or unrecognized instability has been associated with an increased risk of adverse neurodevelopment outcomes and sudden unexpected infant death beyond the immediate NICU stay. However, the review identifies significant knowledge gaps in how to reliably quantify the severity of these events at an individual level [12].

This dissertation’s particular focus is motivated by the fact that this combination—a condition that is prevalent, clinically significant, and challenging to characterize with straightforward criteria is exactly the kind of issue that continuous, high-resolution physiological monitoring combined with machine learning is positioned to address.

2.2 The wider landscape: scope, taxonomy, and recurring limitations

The particular subject of this dissertation is placed within the broader context of AI in neonatal and pediatric care by a number of general reviews. In a scoping review of continuous cardiorespiratory monitoring in hospitalized children in general, Walker et al. screened 2,909 articles and extracted 105 in full, of which 47 described prediction models based on continuous physiological data and 58 described novel monitoring de-

vices or approaches [1]. At the level of the entire field, their key finding—that these studies were “generally small and single center,” that only seven prediction models integrated multidimensional data beyond the physiological signal itself, and that no study took place entirely outside of an intensive care setting establishes precisely the generalizability issue that this dissertation later directly addresses in its own results.

Specifically focusing on wearable neonatal cardiorespiratory monitoring, Sitaula et al. create a taxonomy of machine learning techniques used in this field and report that only 6% of identified studies use transfer learning from pre-trained models, 35% use deep learning, and 63% use traditional machine learning techniques [3]. This dissertation’s decision to compare a classical model (logistic regression) against an ensemble method (random forest) rather than immediately commit to a deep learning architecture given the dataset’s modest scale is directly related to their discussion, which makes clear that traditional techniques continue to be more appropriate for the small datasets typical of this field and offer greater interpretability for clinical trust.

Helm et al. argue that even though the oxycardiorespirogram (OCRG), a related multi-channel monitoring modality that combines respiratory pattern, heart rate variability, transcutaneous oxygen tension, and skin perfusion, is a well-established clinical tool, its rich time-series structure is still largely underutilized for predictive modeling. They specifically call for shallow and deep time-series learning approaches to close this gap [5]. Nikolaou et al. conclude that “data quality, generalizability, interpretability, and limited prospective validation continue to constrain widespread clinical implementation” [6] after taking an even more comprehensive approach to neonatal respiratory care as a whole, including applications to respiratory distress syndrome, bronchopulmonary dysplasia, apnea of prematurity, and ventilatory management.

Rallis et al.’s review of AI across the entire NICU, including neuromonitoring, retinopathy of prematurity, sepsis, and mortality prediction, echoed this conclusion at the broadest scope of all. It also concludes by urging additional retraining, clinical validation, and attention to ethical considerations before these methods see wider adoption [19]. This conclusion’s consistency across four independent reviews covering various sub-domains and publication years is proof in and of itself that the “small, unvali-

dated, single centre” pattern mentioned in Section 2.4 below is a structural feature of the field rather than a singular flaw in any one study.

2.3 Continuous physiological signal detection based methods

The literature that is most directly related to the technical approach of this dissertation comes from a limited group of research that directly apply machine learning to continuous ECG, respiration, and SpO2 signals in order to identify events associated to bradycardia and apnea. Using ECG, chest impedance, and SpO2 signals divided into 30-second intervals with a 5-second shift, Varisco et al. created a pipeline to categorize automatically flagged candidate central apnea events as true or false alarms [9]. In their first analysis, they examined three conventional models: logistic regression with elastic-net regularization, random forest, and support vector machine. These models achieved an AUROC of 0.88–0.90, with much higher correct-detection rates (93–95%) for apneas, particularly bradycardia. Although overall precision dropped significantly outside the training hospital (from 0.54 to 0.28), a follow-up study that expanded this pipeline with convolutional and recurrent neural networks reported bradycardia-paired sensitivity of 0.83–0.88 even on a held-out cohort from a different set of infants [4]. The lack of a publicly available, externally vetted dataset is expressly mentioned as a problem in both studies, and the earlier paper specifically suggests deep learning as a future research topic [9].

Support vector machines, decision trees, naïve Bayes, and KNN classifiers were used in a related but methodologically different study by an unidentified author group to differentiate between true and false cardiorespiratory alarms using heart rate, SpO2, and cerebral oximetry (StO2) in a cohort of 14 infants [14]. This study confirms that the two-model classical comparison used in this dissertation reflects real, active practice in the field rather than an arbitrary simplification, even though it is smaller in scale and does not specifically isolate bradycardia from other alarm types. It also shows that classical, highly interpretable classifiers are still a viable and frequently chosen approach in this literature even outside the Varisco group’s specific pipeline.

2.4 Prediction-based approaches: forecasting ahead of onset

Instead of post-hoc classification of a candidate event that has already been highlighted, a methodologically different and, for the purposes of this research, more immediately applicable line of work focuses on true early-warning prediction prior to event beginning. Joshi et al. achieved an AUROC of up to 0.89 and prevented 26% of critical alarms with a median lead time of 18.4 seconds by using decision-tree and boosted-tree models on a significantly larger cohort (55 infants, over 34,000 monitoring hours) to predict a critical ("red") alarm from the presence of a prior lower-severity ("yellow") alarm [7]. Instead of isolating bradycardia specifically, this method combines multiple alarm kinds, and its predictions are based on the occurrence of a previous lower-severity alarm rather than coming from ongoing background monitoring alone. Nearer to the specific goal of the current study, Zuzarte et al. used a hybrid Gaussian

Table 2.1: Characteristics of the ten PICS recordings, derived from the annotation files. Sampling rates are not uniform across infants.

Infant	ECG (Hz)	Resp. (Hz)	Hours	Onsets	Onsets/h	Median HR (bpm)
1	250	500	45.6	77	1.69	155
2	500	50	43.8	72	1.64	130
3	500	50	43.7	80	1.83	129
4	500	50	46.8	66	1.41	166
5	250	50	48.8	72	1.48	142
6	500	50	48.6	56	1.15	136
7	500	50	20.3	34	1.67	160
8	500	50	24.6	28	1.14	140
9	500	50	70.3	97	1.38	149
10	500	50	47.3	40	0.85	154
Total	–	–	439.8	622	1.41	–

Mixture Model and logistic regression framework [15] to predict apnoea-bradycardia-hypoxaemia (ABH) events prior to onset using features derived from time-delayed correlations of inter-beat and inter-breath intervals, combined with movement features extracted from photoplethysmogram artifact. 75–83% of the most severe bradycardia events were accurately predicted by their model, which had a mean AUC of 0.77. Three

to four minutes prior to onset, half of the accurately predicted events were detected [15]. Crucially, the authors indicate that a statistically significant gain in performance was caused by movement features, and that categorization for most of their training individuals (5 out of 7) was not substantially different from chance when movement features were removed. a conclusion that directly affects this dissertation because there is absolutely no movement or plethysmographic channel in the PICS dataset used here.

2.5 Different demographics, modalities, and definitions of outcomes

Not all of the studied work separates bradycardia prediction from continuous cardiorespiratory waveforms; a number of research use very different inputs to handle closely comparable problems, and their differences are instructive. Using static clinical and demographic information upon admission, such as gestational age, birth weight, Apgar score, and comparable parameters, one study created a risk-prediction model for apnea using LASSO feature selection and logistic regression into a nomogram [13]. A one-time risk score calculated from variables known at a single point in time, as opposed to a continually updating forecast generated from an ongoing physiological signal, constitutes a fundamentally different modeling paradigm from the current dissertation. Instead of providing a clear methodological precedent, it is a helpful contrast that shows how "predicting bradycardia risk" and "predicting the timing of a bradycardia event" are related but different challenges that call for different models and data.

Two other research take a hardware and engineering approach to newborn apnea detection instead of an algorithmic one. One describes a digital hardware (FPGA-based) implementation of a machine-learning model for neonatal sleep apnea detection [8]; another presents a smart neonatal supine-monitoring system that combines vital-sign sensing with thermal regulation, using threshold-based alerting instead of a trained machine learning classifier [2]. Both show that converting any predictive model of the type developed here into deployed clinical hardware is a significant and independent engineering undertaking that should be acknowledged as a boundary of this dissertation's scope, even though neither contributes a modeling technique that is directly

reusable in this dissertation.

A last, more recent study uses a completely different sensing modality: contactless visible-thermal imaging to remotely estimate a breathing signal. One-dimensional convolutional neural networks, LSTM networks, and a hybrid of the two are then used to classify the breathing signal for apnea [16]. Neonatal apnea detection is only mentioned as one of several possible future applications rather than the dataset under examination, and this study is mostly undertaken in a generic (non-neonatal-specific) population. Although it does not directly influence the design choices made in this dissertation, it is included here for the purpose of completing the literature search and because it shows an active alternative direction contact free monitoring—that deviates from the ECG/respiration based method.

2.6 Neonatal AI on a larger scale using a new data modality

Even though they don’t particularly address bradycardia, a different group of research is worth placing next to the ones mentioned above due to the notable differences in dataset size and validation rigor they attain, which are closely related to how this dissertation’s own findings should be interpreted. With an AUROC of 0.861, Kim et al.’s bidirectional LSTM model outperformed random forest, XGBoost, and a traditional early warning score system in predicting the need for invasive mechanical ventilation from electronic health record (EHR) and vital-sign data in 1,394 neonates [17]. Yoneda et al. predicted bronchopulmonary dysplasia and mortality using data available within six hours of delivery using gradient boosting decision trees on a countrywide Japanese registry [18]. Most remarkably, Jang et al. compared seven machine learning models to predict respiratory distress syndrome from perinatal and birth variables using a truly multicenter cohort of 13,087 very low birth weight infants from 76 hospitals in the Korean Neonatal Network. They then proposed a five layer deep neural network ensemble that achieved an AUROC of 0.9187 and was later made available as a public web application [20]. Two key differences between these three studies and the bradycardia-specific literature reviewed above are that they use mostly static, point in time perinatal or EHR variables

instead of continuous physiological waveforms to predict different clinical outcomes, and—more crucially for this dissertation. Compared to the bradycardia-specific studies covered in Sections 2.3 and 2.4, which all comprise cohorts of 10 to 55 infants from a single institution, they operate at a scale that is one to three orders of magnitude larger (1,394 to over 13,000 infants, several spanning dozens of hospitals). These larger studies achieve significantly higher and more stable AUROC values (0.86–0.92) than the bradycardia-specific literature (0.77–0.90, with the lower end of that range coming from research trying actual external validation). This scale mismatch is not accidental. This comparison is used later in the dissertation to contextualize its own findings: based on the evidence of this larger body of literature, it should be expected that a small, single-institution physiological dataset like PICS would be a significantly more difficult setting in which to achieve high discriminative performance than the multicenter, large-N studies achieve for different, more static prediction targets.

2.7 Examined and disregarded studies

Two further studies found during the literature search were eliminated after being evaluated for methodological relevance. One uses a physics-informed autoencoder to extract features from multidimensional gas-sensor measurement data [10], while the other uses a conditional variational autoencoder, tested on urban taxi-demand data [11], to detect anomalies in seasonal time series in real-time. Although neither paper deals with clinical, physiological, or neonatal data of any kind, and no methodological information unique to either paper has been incorporated into the design of this dissertation, both were reviewed on the grounds that their underlying feature-extraction or anomaly detection techniques might generalize to physiological time series. They are mentioned here merely to show that, rather than being ignored, they were taken into consideration and specifically left out due to their relevance.

2.8 Research gap and synthesis

The detection- and prediction-based studies of Sections 2.3 and 2.4 are two limitations that recur consistently throughout the literature most directly relevant to the goal of

this dissertation. These limitations are independently supported by the more comprehensive field-level reviews of Section 2.2. First, every reviewed study that predicts bradycardia or closely related events from continuous physiological signal is based on a small, privately held, single-center cohort of 10 to 55 infants; no external or cross-site validation is reported anywhere in this literature; [9] specifically identifies the lack of a publicly available annotated dataset for this task as a limitation at the time of its publication, and [1]’s field-wide scoping study independently verifies that this pattern small, single-center, rarely externally validated—describes not only this particular sub-literature but the vast majority of continuous physiological monitoring research across hospitalized children in general. Second, movement derived features, which are not always available in routine NICU monitoring and whose absence the same study shows drives performance toward chance for most subjects, were crucial to the strongest predictive performance reported for genuine early warning prediction specifically [15]. By using a similar modeling method to PICS, an independent, publically accessible dataset, this dissertation directly fills the first gap by investigating if claimed bradycardia detection/prediction performance extends beyond the original institution. Additionally, it offers a direct empirical test of the second point: the results of this dissertation provide a real-world test of how much predictive performance is retainable using cardio-respiratory features alone, which is exactly the condition under which [15] reports its own weakest results. This is because PICS only contains ECG and respiration signals, without any movement channel.

Table 2.2: The four studies underpinning this design, and the position of the present work. All prior cohorts are private; PICS is public.

Study	Cohort			Signals	Task	Validation		Headline result
Varisco et al. [9]	et	20 preterm			ECG, CI, SpO ₂	Classify flagged candidate central apnoeas	LOPO, 10-fold	AUROC 0.88 ± 0.08; precision 0.42 at FPR 0.3
Varisco et al. [4]	et	10	+	10	ECG, CI, SpO ₂	Central apnoea detection	LOPO + hold-out	CNN precision 0.54, sensitivity 0.75; 0.87 for apnoeas paired with bradycardia
		preterm, 480 h each						

continued on next page

Table 2.2 – continued from previous page

Study	Cohort	Signals	Task	Validation	Headline result
Zuzarte et al. [15]	10 preterm, 19–72 h	Inter-beat, inter-breath, PPG movement	Pre-event vs inter-event frame	Per-subject k -fold	AUC 0.77 ± 0.04 ; 2.87 false-positive min/h
Joshi et al. [7]	55 infants, >34,000 h	HR, BR, SpO ₂ , HRV	Critical alarm following an advisory alarm	80/20 split	AUROC 0.89, sensitivity 0.33 at specificity 0.98; 18.4 s median lead
This work	10 preterm, 439.8 h, public	ECG R-peaks, respiration peaks	Bradycardia onset within lead H	LOPO	Reported in Chapter 5

Chapter 3

Methodology

The Preterm Infant Cardio-Respiratory Signals (PICS) database, a publicly accessible dataset that includes expert bradycardia onset comments along with continuous ECG and respiration recordings from ten preterm newborns at UMass Memorial Healthcare, is used in this dissertation. Sampling speeds vary each infant and signal type (500 Hz or 250 Hz for ECG; 50 Hz or 500 Hz for respiration), reflecting variations in the particular monitoring equipment used during data collection. Recording lengths range from roughly 20 to 70 hours per infant. The dataset defines a bradycardia event as a heart rate of less than 100 beats per minute (R-R interval more than 0.6 seconds) maintained for at least two consecutive beats. Events that occur within three minutes of one another are combined into a single episode.

Table 3.1: Characteristics of the ten PICS recordings, derived from the annotation files. Sampling rates are not uniform across infants.

Infant	ECG (Hz)	Resp. (Hz)	Hours	Onsets	Onsets/h	Median HR (bpm)
1	250	500	45.6	77	1.69	155
2	500	50	43.8	72	1.64	130
3	500	50	43.7	80	1.83	129
4	500	50	46.8	66	1.41	166
5	250	50	48.8	72	1.48	142
6	500	50	48.6	56	1.15	136
7	500	50	20.3	34	1.67	160
8	500	50	24.6	28	1.14	140
9	500	50	70.3	97	1.38	149
10	500	50	47.3	40	0.85	154
Total	–	–	439.8	622	1.41	–

The methodology proceeds through five stages, summarised below.

Data preparation Along with the bradycardia start times used as ground-truth labels, R-peak locations (manually verified) and respiration-peak locations (algorithmically generated, not manually verified) were extracted from the dataset’s

annotation files. Before modeling, a preliminary evaluation of the data quality was carried out for each of the ten infants, calculating the time of the recording, the derived heart rate, and the derived respiration rate for each infant. This identified two data-quality problems that needed to be addressed explicitly: infant 5’s respiration signal was non-functional for about the first 24 hours of recording, as shown by a nearly total lack of detected respiration peaks during that time; and infant 5’s ECG and respiration recordings ended about an hour apart, necessitating truncation to the overlapping duration. Affected windows were removed from the dataset instead of imputing values over this gap, in accordance with the idea that data shouldn’t be created to fill actual equipment related gaps.

Label definition and windowing The unit of analysis is a four-window set, which consists of four consecutive 30-second windows separated by a 5-second shift and totaling 45 seconds, in accordance with the moving-window scheme of Varisco et al. [9, 4]. Sets were assessed on a 10-second grid for every recording. When a bradycardia onset happens within a lead time of the end of a set’s final window, the set is considered positive. Only stable intervals lasting at least 450 seconds from each annotated onset were used to create negative sets. Zuzarte et al. [15], who define inter-apnoea frames by a minimum 7.5-minute separation from apnea and exclude all intermediate windows from analysis, are followed in this case control construction; Varisco et al. [4] apply the same principle through their concept of distant stable-sets. Instead of being placed in the negative class, sets that were neither positive nor sufficiently distant were eliminated. Since the literature does not specify a single lead time, was considered an experimental variable and assessed at 0, 5, 10, 15, and 20 seconds. For the reasons outlined in Section 4.6, a supplementary condition with the set overlapping the onset - 15 was added only as a pipeline sanity check and is not reported as predictive performance.

Feature extraction Eight features from the R–R interval series (mean, minimum and maximum heart rate, standard deviation, coefficient of variation, RMSSD, linear

trend, beat count) and ten from the inter-breath interval series (mean breathing rate, maximum inter-breath interval, standard deviation, linear trend, breath count, longest pause as a fraction of the window, count of pauses exceeding two seconds, and three edge-sensitive measures described below) were calculated for each 30-second window.

An issue with interval-based summary statistics is addressed by three features. A respiratory pause that is still in progress at the end of a window is not visible to the maximum inter-breath interval because it only records *completed* intervals. To capture open-ended pauses at the window edge, the features `since_breath` (elapsed time from the last detected breath to the window boundary), `ongoing_pause` (the longer of that value and the maximum completed interval), and `since_beat` were added.

These features for all four windows, a 300-second trailing baseline, the difference between the fourth and first windows as a trend descriptor, and the final window’s deviation from that baseline expressed as both a difference and a ratio make up each four-window set, totaling 144 features. The movement-derived features that Zuzarte et al. [15] identify as the primary source of their own predictive gain could not be replicated because PICS lacks a plethysmographic or motion channel.

Model validation and selection A comparison was made between two models. According to Varisco et al. [9], the best-performing classical model was found to be logistic regression with an elastic-net penalty, whose coefficients are log odds ratios and thus directly provide feature relevance. Varisco et al. [4] found that a one-dimensional convolutional network, which consumes the four windows as a length-four sequence of 18 features instead of the flattened vector, performed best on an equivalent four-window-set input. Both made use of balanced class weighting.

Leave-one-patient-out cross-validation was used in all reported evaluations, and only the training infants’ imputation and standardization were fitted within each fold. In contrast to k -fold cross-validation, where windows from the same infant

may appear in both training and test partitions, this protocol adheres to [4, 9] and, as [9] contends, more closely reflects clinical practice.

Evaluation metrics Since a majority-class predictor would score highly with no clinical value under this class distribution, accuracy was excluded. AUROC and average precision were used to report window-level performance. The consecutive-alarm criterion of Varisco et al. [4] was used to report event-level performance: an alarm is triggered only when the model score surpasses the threshold in N consecutive sets. The threshold is a quantile of the stable-period scores of the training infants and is never selected on the test infant. Both [4] and [15] reported false-alarm burden as alarm-active minutes per patient-hour within stable regions, along with event-level precision.

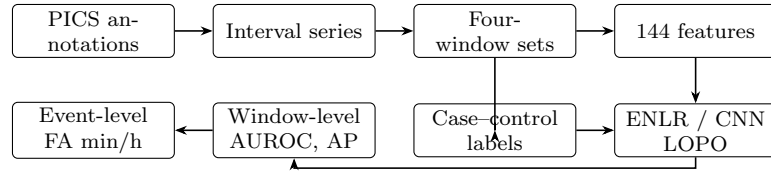


Figure 3.1: Processing pipeline from PICS annotation files to event-level evaluation.

Chapter 4

Experimental Design

4.1 Rationale for the papers underpinning this design

The experimental design was directly informed by four studies chosen from the reviewed literature: [4], [9], [15], and [7]. These papers were chosen not because they discuss neonatal monitoring in general, but rather because they are the only reviewed works that combine three distinct characteristics at the same time: (i) the use of continuous physiological time-series signals that match PICS’s own signal types (respiration and ECG) instead of static clinical variables, registry data, or narrative discussion, (ii) an outcome definition that isolates bradycardia, or an apnea-bradycardia-hypoxaemia pattern in which the severity of bradycardia is reported separately instead of as a pooled or unrelated outcome and (iii) a machine-learning pipeline that has been validated and has quantitative performance metrics that can be compared. Rather from being interchangeable with the others, each paper is also distinct in a way that influenced a particular design choice.

There is a unique pairing between [9] and [4]. The only two reviewed papers that apply the same feature-extraction and windowing pipeline (47 features from 4-window sets of ECG, chest impedance, and SpO2) to the same signals from the same institution are [9], which uses classical machine learning, and [4], which uses convolutional and recurrent neural networks. This dissertation’s own design logic—keeping the feature set, windowing scheme, and validation protocol constant while changing only the model is directly justified by this controlled progression, which is uncommon in the reviewed literature. This way, any performance difference between logistic regression

and random forest can be attributed to model choice rather than being confused by pipeline differences.

[15] is in a unique position on two fronts. First, it is the only reviewed paper that performs actual pre-onset forecasting with bradycardia severity stratified outcomes instead of post-hoc classification of a candidate event that has already been flagged. This aligns with the prediction task that PICS’s own onset time annotations naturally support, making it the closest methodological analog available rather than just a convenient comparison. Second, the description of the PICS dataset itself closely matches its own cited prior work (10 infants, 444 bradycardic events, point-process analysis, AUC 0.79; and a related boosted decision tree result of AUC 0.86), indicating a potential direct empirical benchmark that is not available from the private cohort of any other reviewed paper. Critically, [15] is also particularly significant for interpreting the results of this dissertation rather than just designing them: it reports that, for the majority of its own training subjects, cardio-respiratory features alone, without movement-derived features, produced classification that was not significantly different from chance. This finding directly affects this dissertation because PICS lacks a movement channel.

The scale and evaluation philosophy of [7] are distinct. It is an order of magnitude larger than any other reviewed physiological-signal study, with 55 infants and more than 34,000 monitoring hours. Additionally, it is the only reviewed paper that reports a specific clinical-deployment metric (median lead time in seconds; percentage of alarms preempted) rather than just statistical performance. This influenced the choice to present results in terms of practical lead time rather than just abstract discriminative ability and to report sensitivity and specificity together rather than depending on a single aggregate statistic.

4.2 Exploratory investigation preceding the fixed experimental protocol

An exploratory diagnostic phase was conducted before committing to a defined feature set in order to make sure that subsequent design decisions were supported by data

rather than conjecture. There were three steps to this phase. Initially, a per-infant data-quality audit was carried out for each of the ten infants. This involved calculating the inferred respiratory and heart rates from the raw annotation files and comparing them to physiologically reasonable ranges for premature newborns.

It should be mentioned that there are three simultaneous differences between the first and final iterations: the feature set, the control-sampling strategy, and the lead-time horizon. Therefore, it is impossible to attribute the improvement in performance between them to a single change, and Table 4.1 is reported as a documented negative control instead of a controlled ablation. It was not possible to vary each factor separately within the project timeframe in order to isolate its contribution.

An approximately one-hour difference between infant 5’s ECG and respiration recording durations and a persistent lack of detected respiration peaks for approximately the first 24 hours of recording were the two real equipment-related anomalies that this audit revealed. Both of these anomalies were later verified, rather than presumed, by directly examining the timing and distribution of the impacted windows prior to any exclusion decisions being made.

Second, R-R interval trajectories in the 300 seconds previous to several distinct bradycardia onsets across multiple infants were plotted and visually examined before the feature set was finalized. This examination showed a heterogeneous outcome consistent with the subject level variability reported in [15], with some onsets showing a late, abrupt rise in R-R interval in the final 15 to 50 seconds prior to onset and others showing no such pattern. Two subsequent design modifications were directly inspired by this evidence: lowering the window step size from 30 to 10 seconds to increase the probability that a particular window would capture this narrow pre-onset transition, and adding a trend feature that could differentiate between a rising R-R interval sequence and a flat one, which basic summary statistics like the mean.

Third, before the final design was fixed, a full first iteration of the modeling protocol was run and assessed. This version employed lead-time horizons of 1, 3, and 5 minutes, labeled each window throughout the entire recordings instead of sampling controls from far-off stable periods, and combined a linear trend term with a reduced feature

set of aggregate interval statistics. Three model families logistic regression with elastic-net regularization, random forest, and a one-dimensional convolutional network were assessed using leave-one-patient-out cross-validation. Table 4.1 presents the findings. Each model and horizon combination produced a mean AUROC that was statistically indistinguishable from chance, ranging from 0.506 to 0.521. This result’s consistency across three structurally different model families is instructive in and of itself: a linear model, a bagged tree ensemble, and a convolutional network all share nearly no inductive bias, so their identical failure suggests that the experimental design, not the model’s capacity, is the limiting factor. Inadequate model flexibility would have been a reasonable explanation if only the linear model had failed, but it isn’t available in this case.

This result inspired the lead-time analysis of Figure 4.1, which localized the available predictive information to within about twenty seconds of onset. It also led to two design modifications that were incorporated into the final protocol: the lead-time horizons were shortened from minutes to seconds, and the case-control sampling scheme of Zuzarte et al. [15] was used instead of labeling each window in the recording.

Table 4.1: First-iteration results, retained as a negative control. Mean AUROC ($\pm$ standard deviation) across leave-one-patient-out folds. No combination separates from chance. Positive-window rates were 2.44 %, 7.30 % and 12.16 % at the 1, 3 and 5-minute horizons respectively, across 142,500 retained windows.

Model	1 min	3 min	5 min
Elastic-net logistic regression	0.521 $\pm$ 0.061	0.512 $\pm$ 0.027	0.514 $\pm$ 0.021
Random forest	0.517 $\pm$ 0.069	0.513 $\pm$ 0.025	0.516 $\pm$ 0.020
1D convolutional network	0.517 $\pm$ 0.059	0.506 $\pm$ 0.040	0.508 $\pm$ 0.023

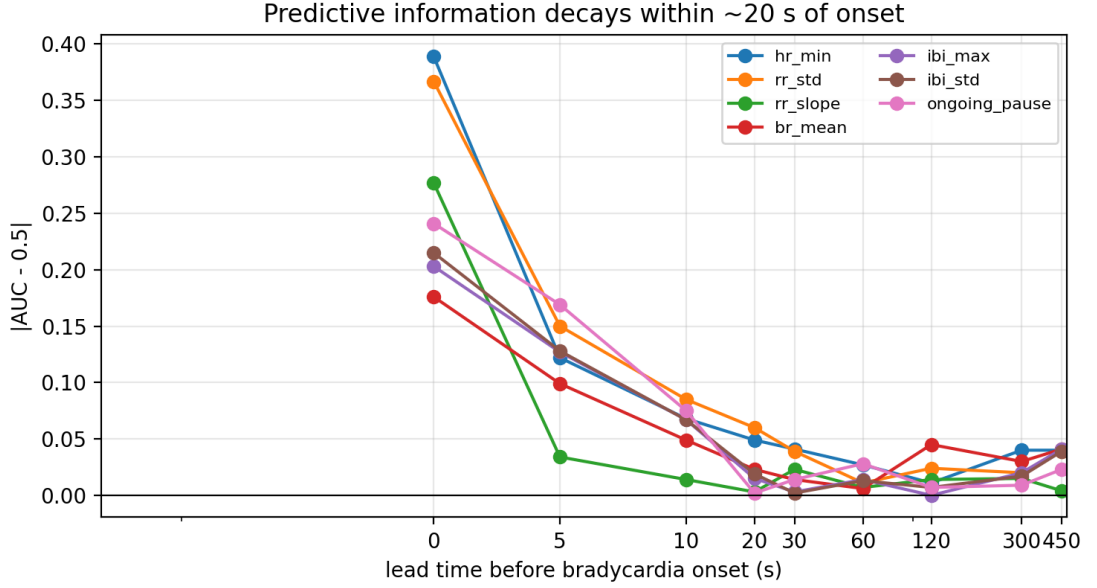


Figure 4.1: Univariate discriminative information as a function of lead time before bradycardia onset, averaged across the ten infants. Distance from 0.5 is plotted so that features separating in either direction are comparable. Information is concentrated within roughly 20 seconds of onset and is indistinguishable from chance beyond 60 seconds.

4.3 Experimental variables

- **Independent variables:** prediction lead time ($H \in \{0, 5, 10, 15, 20\}$ s); signal set (all features, respiration only, cardiac only); model type (elastic-net logistic regression, one-dimensional CNN).
- **Dependent variables:** sensitivity, accuracy, and false alarm minutes per patient-hour at the event level; AUROC and average precision at the window level.
- **Controlled variables:** window length (30,s), four-window set structure, 450,s control guard, cross-validation protocol, and class weighting were all kept constant for each combination.

4.4 Hypotheses

- H1.** Beyond a lead of a few tens of seconds, discriminative performance approaches chance due to the rapid decay of predictive information about an impending bradycardia. Joshi et al. [7] reported a median lead time of 18.4 s and noted that performance improves with a shorter prediction window. In this work, it is tested twice: multivariately by the first-iteration result of Table 4.1, and univariately in Figure 4.1.
- H2.** Given that central apnea precedes the associated bradycardia and that respiration features are ranked highest in relevance by Varisco et al. [9], respiration-derived features carry more predictive information than heart-rate-derived features at non-zero lead times.
- H3.** Because PICS does not supply SpO_2 or the movement channel that [15] identifies as the source of its own predictive gain, absolute performance will be lower than the AUROC values reported in [4, 9, 15].
- H4.** Compared to the large- N neonatal AI studies of Section 2.6 [17, 18, 20], performance will be lower on average and more variable across leave-one-patient-out folds, which is consistent with the field-wide pattern found by [1].

4.5 Justification of key design decisions

Table 4.2: Key experimental design decisions and their basis in the literature.

Decision	Basis	Justification
Four 30 s windows, 5 s shift, as one record	[9, 4]	Directly adopted; [9] report AUROC rising from 0.73 to 0.88 as windows increase from one to four

continued on next page

Table 4.2 – continued from previous page

Decision	Basis	Justification
Scoring grid stride of 10 s	–	Experimental compromise for computational tractability; disclosed as such rather than presented as a literature-matched choice
Lead times 0–20 s	[7]	[7] report a median lead of 18.4 s; longer horizons were rejected on the evidence of the exploratory analysis in Section 4.2
Controls ≥ 450 s from any onset	[15, 4]	[15] define inter-apnoea frames by a 7.5-minute minimum separation; [4] use distant stable-sets equivalently
ENLR and 1D CNN	[9, 4]	[9]’s best classical model against [4]’s best neural architecture, with feature set and validation held constant
Leave-one-patient-out CV	[4, 9]	Prevents subject-level leakage and reflects deployment on an unseen infant
Consecutive-alarm criterion	[4]	The principal mechanism by which [4] trade sensitivity for precision

Beyond the preceding synopsis, two choices warrant a more thorough examination of the options that were taken into consideration and rejected. In terms of window step size, a step that matched the 5-second shift of [9] and [4] was taken into consideration but rejected in favor of 10 seconds because the resulting window count under a 5-second step would have tripled compute time for the subsequent hyperparameter tuning and leave one infant out evaluation stages beyond what was feasible within the available project timeline; this trade-off is disclosed explicitly here rather than presented as a literature matched choice. In terms of model selection, a deep learning architecture that matched [4] was taken into consideration but ultimately rejected in favor of two

classical models. This is because [4]’s own results demonstrate that a deep learning approach necessitates a larger and more carefully curated dataset than PICS provides in order to prevent overfitting, and [3]’s review explicitly notes that classical methods remain better suited to small datasets typical of this field.

4.6 Threats to validity

The validity of this experimental design is expressly threatened by a number of factors. In terms of internal validity, the main risk is data leakage between training and test data. This risk is reduced by fitting all preprocessing and hyperparameter selection solely on training data within each fold, using leave one infant out cross-validation for final evaluation, and making sure no infant’s windows appear in both training and test sets in any fold. In terms of external validity, any performance obtained cannot be assumed to generalize to infants monitored with different equipment, at different institutions, or under different clinical protocols due to the dataset’s small size (10 infants) and single-institution origin—a limitation shared with [4], [9], and [15], and identified at the level of the wider field by [1]. In terms of construct validity, the three parallel horizons tested here should be interpreted as an exploration of this design space rather than a claim that any one horizon is the clinically correct choice. The definition of a ”successful prediction” (a bradycardia onset occurring within a fixed lead-time window following a given signal window) is one reasonable operationalization of early warning, but it is not the only one possible.

Another risk is unique to the outcome definition employed here and relates to construct validity. Since bradycardia is defined *from* heart rate, any heart-rate-derived feature measured at or right before an annotated onset is partially circular: rather than predicting the event, the model may be witnessing the start of the very deceleration that makes up the event. This is addressed by two actions. First, the situation where the four-window set overlaps the onset is never reported as predictive performance; rather, it is only reported as a pipeline sanity check. Second, a respiration-only model was assessed throughout; any discrimination it achieves cannot result from this circularity since respiration timing is an independent channel.

Rather than the model, the evaluation poses a second risk. Regardless of whether alarms are timed to events or not, the frequency of an alarm’s activity can inflate event-level sensitivity measured over a detection window. The proper control is a permutation test that preserves alarm rate and clustering while destroying their temporal relationship by shifting the alarm series in a circular fashion with respect to the actual onsets. This control is listed in Section 6.6 as the most significant outstanding validation of the event-level results reported in Chapter 5, but it was not completed within the project’s timeframe.

4.7 Tools and computational environment

Using the `wfdb` library to read the native WFDB-format signal and annotation files from the PICS dataset, `numpy` and `pandas` for handling numerical and tabular data, and `scikit-learn` for model implementation, cross-validation, and evaluation metrics, all data processing, feature extraction, and modeling were done in Python. To enable the tiered, exploratory research outlined in Section 4.2, development was carried out interactively in a Jupyter notebook environment.

Chapter 5

Results

This chapter reports the measured performance of the final protocol described in Chapter 3. Results are presented at two levels. Window level results (Sections 5.1–5.3) treat each four-window set as an independent record and answer whether the models discriminate at all. Event level results (Sections 5.4–5.5) treat each annotated bradycardia as the unit of analysis and answer whether the model would function as an alarm. The two are reported separately because they are not interchangeable: strong window-level discrimination does not guarantee usable alarm behaviour under continuous monitoring, and Section 5.5 shows that in this case it does not.

5.1 Window level performance by lead time and signal set

For each lead time and each of the three signal sets all 144 features, respiration-derived features only, and cardiac derived features only Table 5.1 displays the mean AUROC and average precision across the ten leave-one-patient-out folds.

Here are three observations. First, lead time causes discriminative performance to decline sharply. The complete feature set achieves an AUROC of 0.922 at a lead of zero seconds, where the last window ends precisely at the annotated onset. This has decreased by five seconds to 0.773, ten seconds to 0.627, and fifteen seconds to 0.537. The model is indistinguishable from chance at twenty seconds. This supports H1 and is in line with both the first-iteration result of Table 4.1 and the univariate analysis of Figure 4.1.

Second, there is no performance outcome for the row at a lead of -15 seconds. The heart-rate features in this condition observe the bradycardia itself because the four window set overlaps the annotated onset; the AUROC of 0.995 is therefore tautological

and is only reported to verify that the pipeline detects the events it is given. It is not included in any further analysis or comparisons with earlier research.

Third, and perhaps most importantly, lead time causes the two signal sets' rankings to reverse. Given that bradycardia is defined by heart rate, the cardiac features are the stronger set at a lead of zero seconds (0.910 against 0.821). This is reversed at a five second lead: the respiration only model performs nearly identically to the full feature set, and respiration derived features achieve an AUROC of 0.764 while cardiac derived features only reach 0.619. For the reasons listed in Section 5.3, this is the main finding of this dissertation and supports H2.

Table 5.1: Window-level performance of the elastic-net model. Mean AUROC across leave-one-patient-out folds, with average precision in parentheses. The -15 s row is a pipeline sanity check, not a performance result.

Lead (s)	All features	Respiration only	Cardiac only
-15	0.995 (0.907)	0.838 (0.318)	0.995 (0.904)
0	0.922 (0.531)	0.821 (0.257)	0.910 (0.506)
5	0.773 (0.182)	0.764 (0.165)	0.619 (0.097)
10	0.627 (0.086)	0.630 (0.085)	0.546 (0.060)
15	0.537 (0.055)	0.538 (0.055)	0.528 (0.051)
20	0.518 (0.047)	0.525 (0.051)	0.502 (0.045)

5.2 Comparison of the two models

The elastic-net logistic regression and the one-dimensional convolutional network on the entire feature set at the three lead times where a signal is present are compared in Table 5.2.

Every time a prediction is made, the linear model outperforms the other models by 0.068 AUROC at five seconds and 0.058 at ten seconds. The two are nearly identical at a lead of zero seconds, with the convolutional network slightly lower on AUROC (0.909 versus 0.922) but higher on average precision (0.558 versus 0.531), suggesting

a better ranking of the positive class at the operating region that is most important under class imbalance.

This does not replicate the ordering reported by Varisco et al. [4], who discovered that on an equivalent four-window-set input, convolutional networks performed better than elastic-net regression. There are two pertinent distinctions. Their feature set was derived from three signals, including SpO₂ and chest impedance waveforms, instead of the two available in PICS. Their reference dataset included 951 annotated central apneas against the 622 bradycardia onsets available here. The most likely reason for the reversal is the smaller, lower-dimensional training set used here, as a convolutional architecture has more parameters to fit than a penalized linear model. This result does not support a general superiority claim for either model.

Table 5.2: Model comparison on the full feature set. Mean AUROC across leave-one-patient-out folds, average precision in parentheses.

Lead (s)	Elastic-net LR	1D CNN
0	0.922 (0.531)	0.909 (0.558)
5	0.773 (0.182)	0.705 (0.123)
10	0.627 (0.086)	0.569 (0.066)

5.3 Feature relevance

For the elastic-net model at a five-second lead, Figure 5.1 displays the median log odds ratio of the twenty most significant features across the leave-one-patient-out folds. This analysis does not require a post-hoc attribution method because the coefficients of a penalized logistic regression are log odds ratios; Varisco et al. [9] used the same method to rank feature relevance in their own model.

The three most important characteristics are all related to breathing: the ratio of the continuous respiratory pause to the infant’s own trailing baseline, the change in mean breathing rate, and the change in breath count throughout the four-window set. Sixth is the first heart-rate feature. This is in line with [9], whose log odds ratio analysis also gave respiration features a higher ranking than cardiac, SpO₂, and signal-quality features.

Beyond feature ranking, two aspects of this outcome are significant. This approach

differs from the threshold-based alarming described in [2] because the dominant features are all *change* features rather than absolute levels, indicating that the model is responding to a departure from the infant’s own recent respiratory state rather than to an absolute threshold. Separately, two of the ten most important characteristics are `ongoing_pause` and `since_breath`. The feature design choice outlined in Chapter 3 is supported by the fact that neither can be calculated from completed inter-breath intervals alone.

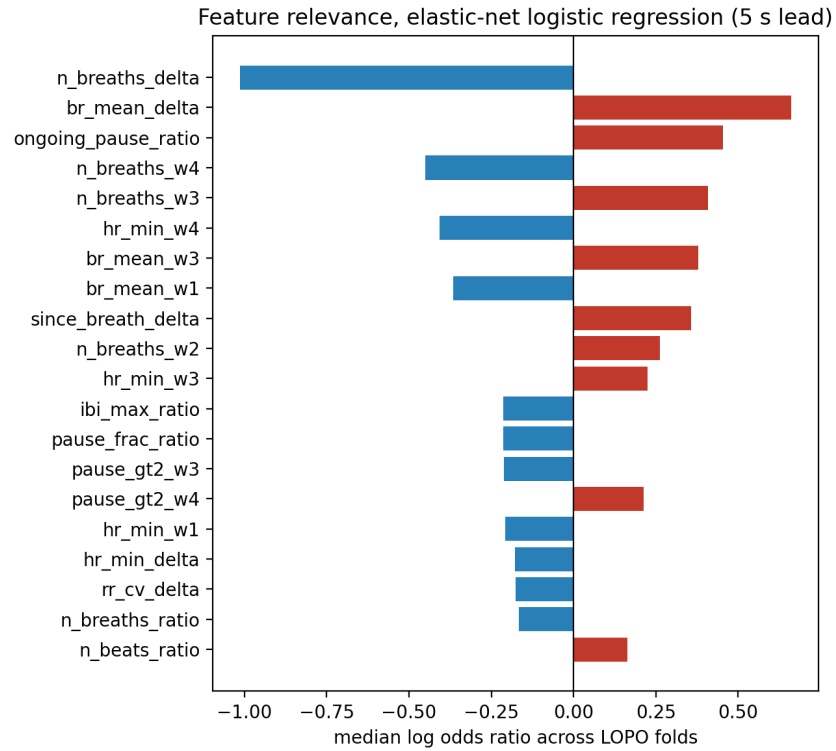


Figure 5.1: Median log odds ratios across leave one patient out folds for the twenty most influential features of the elastic-net model at a five-second lead. The three most influential features are respiratory.

5.4 Performance at the event level

The model’s behavior as an alarm is not described by window-level AUROC. Thus, under the consecutive-alarm criterion of Varisco et al. [4], Table 5.3 presents event-level results at two detection windows and four alarm thresholds, with false-alarm burden expressed in alarm active minutes per patient-hour.

Burden and performance trade in the anticipated direction. At a cost of 6.40 alarm-minutes per patient-hour, the model detects 54.4% of bradycardia events within thirty seconds of onset at its most permissive operating point. Requiring three consecutive positive sets lowers the sensitivity to 44.1% and the false alarm burden to 2.64 minutes per patient hour. The burden is further reduced to 1.38 minutes per patient hour at a sensitivity of 29.5%, when the threshold is raised to the 98th percentile of stable period scores.

Widening the detection window from thirty to sixty seconds increases sensitivity by five to seven percentage points at the same false alarm burden. This is because the criterion for crediting a detection is the only thing that changes, and the alarm series remains the same.

At the event level, the respiration-only model performs worse than the window-level results of Table 5.1, with a sensitivity of 47.7 % at 6.13 minutes per patient-hour compared to 54.4 % for the full model at a comparable burden. As a result, the respiration-only ablation continues to serve as the foundation for the interpretation in Section 5.3, while the complete feature set is kept for the operating characteristics described here.

Table 5.3: Event-level performance, leave-one-patient-out. Threshold is a quantile of the training infants' stable-period scores and is never selected on the test infant. FA denotes alarm-active minutes per patient-hour in stable regions.

Features	Det. win.	Threshold	N consec.	Sens. / FA (min/h)
All	30 s	0.90	1	0.544 / 6.40
All	30 s	0.90	3	0.441 / 2.64
All	30 s	0.95	1	0.409 / 3.30
All	30 s	0.95	3	0.303 / 1.16
All	30 s	0.98	1	0.295 / 1.38
All	30 s	0.99	1	0.216 / 0.81
All	60 s	0.90	1	0.611 / 6.40
All	60 s	0.90	3	0.458 / 2.64
All	60 s	0.95	1	0.461 / 3.30

continued on next page

Table 5.3 – continued from previous page

Features	Det. win.	Threshold	N consec.	Sens. / FA (min/h)
All	60 s	0.98	1	0.320 / 1.38
Resp. only	30 s	0.90	1	0.477 / 6.13
Resp. only	30 s	0.95	1	0.344 / 2.96
Resp. only	60 s	0.90	1	0.566 / 6.13
Resp. only	60 s	0.95	1	0.413 / 2.96

5.5 Inter-patient variation and event-level precision

The alarm burden a clinician would encounter is not adequately described by sensitivity and false-alarm burden alone. Therefore, at the operating point of the 90th-percentile threshold with three consecutive sets and a thirty second detection window, Table 5.4 reports, per infant, the number of discrete alarm episodes raised, how many of those preceded a genuine bradycardia, and the resulting precision.

With a precision of 0.124 at a pooled sensitivity of 0.428 and a mean alarm rate of 5.35 per hour, the model raised 2,216 alarm episodes across the cohort, of which 274 preceded an annotated bradycardia. To put it simply, this model would raise about seven out of every eight false alarms.

The original project proposal’s goal of exceeding the precision of 0.35 reported for traditional threshold based alarming [4, 9] is not met by this. The current figure describes bradycardia prediction at a thirty second lead under an alarm episode definition established in this work, while the 0.35 figure describes central apnea alarms generated by chest impedance monitors and scored against expert apnea annotations. The denominator, the signal set, and the event are all different. However, no operating point in Table 5.3 would satisfy the original goal.

Infants’ performances also differ significantly from one another. The precision ranges from 0.049 to 0.220, the factor is 5.5, and the alarm rate is between 1.88 and 10.32 per hour. Across the cohort, sensitivity and alarm rate have a strong correlation ($r = 0.74$), suggesting that a large portion of the variation in sensitivity between infants is due

to how frequently the model fires in that infant rather than how well it discriminates there. This shows that a single population threshold does not apply to an unseen infant and that per-infant calibration would be necessary for deployment because the alarm threshold is set as a quantile of the *training* infants’ score distribution.

Additionally, infants with more frequent bradycardias have higher precision ($r = 0.65$, $n = 10$). The same mechanism is reported by Varisco et al. [4], who attribute their own precision decline from 0.54 to 0.28 on a hold-out cohort to “the inclusion of preterm infants with far fewer CAs relative to rejections, which impacts the true positives in the precision calculation.” This is a trend rather than a proven relationship with ten infants, but it replicates a known effect on a separate dataset.

At 0.167 sensitivity and 0.080 precision, infant 5 is the weakest case in the cohort. Additionally, this baby’s respiration annotations are the sparsest, with 0.47 detected breath peaks per second compared to a cohort median close to 0.94. This data-quality issue was discovered prior to modeling and is detailed in Chapter 3. The most likely explanation for this infant’s performance is degraded respiration annotation, since respiratory features are the most significant in this model (Section 5.3).

Table 5.4: Event-level precision per infant at the 90th-percentile threshold with three consecutive sets and a thirty-second detection window.

Infant	Events	Alarm ep.	True	Precision	Sens.	Alarms/h
1	77	135	29	0.215	0.364	2.96
2	72	314	42	0.134	0.542	7.16
3	80	329	50	0.152	0.600	7.53
4	66	88	15	0.170	0.227	1.88
5	72	150	12	0.080	0.167	3.08
6	56	186	18	0.097	0.321	3.83
7	34	97	21	0.216	0.559	4.77
8	28	224	11	0.049	0.393	9.10
9	97	205	45	0.220	0.464	2.92
10	40	488	31	0.064	0.775	10.32
Pooled	622	2216	274	0.124	0.428	5.35

5.6 Summary of hypothesis outcomes

- H1 — supported.** Lead time causes predictive information to rapidly deteriorate. At a fifteen-second lead, AUROC drops from 0.922 to 0.537, and at twenty seconds, it is indistinguishable from chance. Both the first-iteration result of Table 4.1 and the univariate analysis of Figure 4.1 independently lead to the same conclusion.
- H2 — supported.** The three most important features in the model are respiratory, with respiration-derived features reaching an AUROC of 0.764 at a five-second lead compared to 0.619 for cardiac-derived features.
- H3 — supported.** The event-level precision of 0.124 is below the 0.35 threshold-alarming baseline that the project aimed to surpass, and performance is below the values reported in [4, 9, 15].
- H4 — supported.** Performance varies significantly across leave-one-patient-out folds, with alarm rates ranging from a factor of 5.5 and per-infant precision ranging from 0.049 to 0.220. This is consistent with the field-wide pattern of small single-center cohorts found by [1].

Chapter 6

Discussion

6.1 Principal findings

Chapter 5 yields three conclusions. First, the dataset’s predictive information about an impending bradycardia is concentrated within twenty seconds of onset and becomes indistinguishable from chance after that. Second, respiration timing provides more predictive information than heart rate within that limited window. Third, the final model achieves 0.124 at a pooled sensitivity of 0.428, falling short of an alarm precision that would be clinically useful.

Contributions are the first two. This chapter contends that the third result, which is negative, is instructive rather than just a deficiency.

6.2 The lead time horizon

The rest of the work is based on the measurement, not the assumption, of the decay of predictive information with lead time (Figure 4.1, Table 5.1). Three separate methods were used to establish it: univariately, using per-feature discrimination against distant controls; multivariately, using the near-chance first iteration across three model families (Table 4.1); and independently, using the final protocol’s own decay from 0.922 at onset to 0.518 at a twenty-second lead.

This is in order of magnitude consistent with the only previous study that explicitly reported a lead time in the reviewed literature. Joshi et al. [7] noted that ”predictions further into the future are more challenging” and found that performance improved as the prediction window was shortened. They predicted critical alarms a median of 18.4 seconds in advance. Using a publicly available dataset, the current result provides a quantitative account of how quickly that difficulty increases for bradycardia

in particular.

As a practical result, early warning bradycardia prediction based solely on ECG and respiration derived from impedance provides a window of seconds rather than minutes. Until an explicit lead time analysis is provided, any claim of minute scale warning from these two signals should be regarded with caution.

6.3 Respiration against heart rate

The three most important features in the model were respiratory (Figure 5.1), with respiration-derived features reaching an AUROC of 0.764 and cardiac-derived features reaching 0.619 at a five-second lead. For two reasons, this outcome is more significant than its size would indicate.

Firstly, it's not circular. Since bradycardia is determined by heart rate, any cardiac characteristic that is measured near the onset may be witnessing the start of the event's deceleration rather than predicting it. Since respiration is a separate channel, discrimination cannot be accomplished through it alone. In order to provide a measurement free of this confounding factor, the respiration-only condition was added.

The second is that it can be understood physiologically. Premature apnea is caused by an immature central breathing control system, and the accompanying bradycardia occurs after the respiratory pause rather than before it [12]. Therefore, a model is acting in accordance with the mechanism rather than just taking advantage of a statistical artifact if it finds respiratory change to be more informative than cardiac change at a short positive lead. Additionally, it is consistent with the feature relevance reported by Varisco et al. [9], whose log odds ratio analysis prioritized respiration features over cardiac and SpO₂ features in their own model.

Changes rather than levels were the most important features: the ongoing respiratory pause expressed in relation to the infant's own trailing baseline, the change in mean breathing rate, and the change in breath count across the window set. The fixed-threshold alerting described in [2], which compares absolute values against a single population cut-off and cannot represent departure from an individual baseline, is significantly different from this.

6.4 Comparison with previous work

There is no scientifically sound numerical comparison with the reviewed studies, and none is provided. It is worthwhile to express the reasons clearly.

Varisco et al. [9, 4] use ECG, chest impedance, and SpO₂ to categorize candidate central apnea events that a monitor has already detected. While the current task is forecasting from continuous background monitoring with a different outcome, a different signal set, and a different denominator for precision, their task is post-hoc adjudication of a flagged event. Joshi et al. [7] limit prediction to periods of current physiological instability rather than operating continuously by conditioning a critical alarm on an earlier advisory alarm. The unit of analysis for Zuzarte et al. [15] is a 7.5-minute pre-event frame compared to an inter-event frame, which is two orders of magnitude longer than the 45-second window set used here. The strongest models use movement features from photoplethysmograms that are absent from PICS.

The qualitative pattern is comparable. In five of their seven training subjects, Zuzarte et al. report that classification using cardio-respiratory features alone, without movement-derived features, was "not significantly different from chance" [15]. Their findings on an independent dataset are consistent with the near-chance first iteration reported in Table 4.1, across three model families on precisely that signal configuration. Therefore, rather than contradicting the published literature, this dissertation's weakest result supports it.

6.5 Alarm burden and clinical implications

Approximately seven out of every eight alarms would be false at the operating point shown in Table 5.4. No operating point in Table 5.3 would satisfy the project proposal's goal, which was to surpass the precision of 0.35 attributed to traditional threshold-based alarming [4, 9]. That goal was set before the lead-time characteristics of this dataset were known, and Chapter 5 demonstrates that it could not be achieved with the two signals at hand.

Deployment is impacted by two secondary findings. A threshold calibrated on a pop-

ulation of training infants does not transfer to an unseen infant, as evidenced by the $r = 0.74$ correlation between sensitivity and alarm rate across infants: the same threshold generates 1.88 alarms per hour in one infant and 10.32 in another. Instead of using a fixed population threshold, any clinical implementation would need per-infant calibration on an initial monitoring period.

Additionally, infants with more frequent bradycardias had higher precision ($r = 0.65$, $n = 10$). When their precision dropped from 0.54 to 0.28 on a hold-out cohort, Varisco et al. [4] report the same mechanism, attributing it to the inclusion of infants with significantly fewer events compared to rejections. This is a trend rather than a proven relationship with ten infants, but it replicates a documented effect on a separate dataset and implies that reported precision figures in this literature are sensitive to the test cohort's event prevalence.

6.6 Limitations

- **No permutation control on the event level results.** Alarm density alone has the potential to inflate event level sensitivity measured over a detection window. The event level figures in Table 5.3 and Table 5.4 should be interpreted as uncontrolled for this effect because the circular shift control mentioned in Section 4.6 was not carried out. This is the work's most significant exceptional validation.
- **Cohort size and origin.** Ten infants without any outside validation from a single database. This is the limitation shared by [4, 9, 15] and found throughout the entire reviewed literature [1].
- **Reduced signal set.** The three signal feature sets of [9, 4] and the movement features of [15] could not be replicated because PICS lacks both a SpO₂ channel and a plethysmographic or motion channel.
- **Annotation structure.** Only bradycardia onset times without offsets are provided by the database. The severity stratified analysis reported by [15] could not be repeated because event duration and severity could not be inferred from the

annotations.

- **Respiration annotation quality.** Infant 5’s annotation density is about half the cohort median, and respiration peaks are identified algorithmically rather than by hand. This is the most likely explanation for that baby’s subpar performance because respiratory characteristics are the most significant.
- **Uncontrolled comparison between iterations.** The improvement between the first and final protocols cannot be attributed to a single change because they differ simultaneously in lead time, control sampling, and feature set.
- **Circularity at zero lead.** For cardiac features, results at a lead of zero seconds are still somewhat circular. This issue is lessened but not completely eliminated by the respiration only condition.

6.7 Future work

The permutation control mentioned above is the top priority right now. It doesn’t require any additional data and would determine how much of the reported event-level sensitivity is higher than what alarm density alone would yield.

Beyond that, the results lead to three directions. The first directly valid numerical comparison with a published result on this task would be possible with a frame-level replication of the design employed by Zuzarte et al. [15], which classifies 7.5-minute pre-event frames against inter-event frames. Instead of using the photoplethysmographic movement estimate that is not available in PICS, a motion proxy derived from the ECG would test whether the movement contribution that [15] identify as decisive can be recovered from a signal that this dataset does contain. Lastly, the threshold-transfer issue mentioned in Section 6.5 would be resolved by per-infant threshold calibration on an initial monitoring segment, which is a requirement for any deployment.

6.8 Conclusion

This dissertation used the windowing, feature, and validation design developed by Varisco et al. [9, 4] to predict bradycardia on the public PICS database, a setting where

no such evaluation has been published before. It found that predictive information is limited to about the last twenty seconds prior to onset by measuring, rather than assuming, the lead time over which prediction is possible from ECG and impedance derived respiration alone. Respiratory change in relation to an infant's own baseline was identified as the dominant feature family within that window, demonstrating that respiration timing is more informative than heart rate result that is both non circular and physiologically consistent.

Additionally, it was discovered that the goal of surpassing the 0.35 threshold-alarming baseline was not achieved and that the resulting model does not attain a clinically useful alarm precision. This result is in line with the clearly stated conclusion of Zuzarte et al. [15] that cardio-respiratory characteristics by themselves, in the absence of movement information, do not consistently outperform chance. Therefore, this work's contribution is a quantified characterization of what can and cannot be accomplished from the reduced signal set that a public dataset makes available, evaluated under leave-one-patient-out cross-validation, and reported without exaggeration.

References

- [1] S. B. Walker et al., “Novel approaches to capturing and using continuous cardiorespiratory physiological data in hospitalized children,” *Pediatric Research*, vol. 93, no. 2, pp. 396–404, Jan. 2023, doi: 10.1038/s41390-022-02359-3.
- [2] K. Shanmugapriya, S. Mangai, K. R. Laksita, S. Muthukumari, V. Sowmiya, and M. Subhasankari, “Smart Neonatal Supine Monitoring System with Vital Signs and Thermal Regulation,” *Advances in Engineering Research*, pp. 851–862, 2026, doi: 10.2991/978-94-6239-654-8_66.
- [3] C. Sitaula, E. Grooby, T. C. Kwok, D. Sharkey, F. Marzbanrad, and A. Malhotra, “Artificial intelligence-driven wearable technologies for neonatal cardiorespiratory monitoring. Part 2: artificial intelligence,” *Pediatric Research*, vol. 93, no. 2, pp. 426–436, Dec. 2022, doi: 10.1038/s41390-022-02417-w.
- [4] G. Varisco, Z. Peng, H. Niemarkt, P. Andriessen, C. van Pul, and X. Long, “Improving central apnea detection in preterm infants: Neural network strategies and clinical impact,” *Computer Methods and Programs in Biomedicine*, vol. 272, p. 109038, Dec. 2025, doi: 10.1016/j.cmpb.2025.109038.
- [5] K. Helm et al., “Oxycardiorespirograms in neonatal monitoring—current gaps and future potential of artificial intelligence: a mini-review,” *Frontiers in Pediatrics*, vol. 13, Oct. 2025, doi: 10.3389/fped.2025.1680074.
- [6] A. Nikolaou et al., “Artificial Intelligence in Neonatal Respiratory Care: Current Applications and Future Directions,” *Applied Sciences*, vol. 16, no. 3, p. 1339, Jan. 2026, doi: 10.3390/app16031339.
- [7] R. Joshi, Z. Peng, X. Long, L. Feijs, P. Andriessen, and C. Van Pul, “Predictive Monitoring of Critical Cardiorespiratory Alarms in Neonates Under Intensive Care,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–10, 2019, doi: 10.1109/JTEHM.2019.2953520.
- [8] O. Hassan, D. Parvin, and Syed Kamrul, “Machine Learning Model Based Digital Hardware System Design for Detection of Sleep Apnea Among Neonatal Infants,” pp. 607–610, Aug. 2020, doi: 10.1109/mwscas48704.2020.9184554.
- [9] G. Varisco et al., “Central apnea detection in premature infants using machine learning,” *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107155, Nov. 2022, doi: 10.1016/j.cmpb.2022.107155.
- [10] C. Gentili, C. Hierold, and C. I. Roman, “Physics-informed autoencoder for feature extraction from multidimensional sensor measurement data,” *Applied Intelligence*, vol. 55, no. 16, Oct. 2025, doi: 10.1007/s10489-025-06892-6..

- [11] L. Porcelli, M. Trovati, and F. Palmieri, “Real-time anomaly detection in seasonal time series with conditional variational autoencoder,” *Applied Soft Computing*, vol. 184, p. 113761, Aug. 2025, doi: 10.1016/j.asoc.2025.113761.
- [12] G. Erickson, N. R. Dobson, and C. E. Hunt, “Immature control of breathing and apnea of prematurity: the known and unknown,” *Journal of Perinatology*, vol. 41, no. 9, pp. 2111–2123, Mar. 2021, doi: 10.1038/s41372-021-01010-z.
- [13] X. Xu, L. Li, D. Chen, S. Chen, L. Chen, and X. Feng, “Establishment and validation of apnea risk prediction models in preterm infants: a retrospective case control study,” *BMC Pediatrics*, vol. 24, no. 1, Oct. 2024, doi: 10.1186/s12887-024-05125-y.
- [14] D. Ostojic et al., “Reducing False Alarm Rates in Neonatal Intensive Care: A New Machine Learning Approach,” *Advances in experimental medicine and biology* (Print), pp. 285–290, Jan. 2020, doi: 10.1007/978-3-030-34461-0_36.
- [15] I. Zuzarte, D. Sternad, and D. Paydarfar, “Predicting apneic events in preterm infants using cardio-respiratory and movement features,” *Computer Methods and Programs in Biomedicine*, vol. 209, p. 106321, Sept. 2021, doi: 10.1016/j.cmpb.2021.106321.
- [16] L. Maurya, R. Zwiggelaar, D. Chawla, and P. Mahapatra, “Contactless apnea event detection using visible-thermal imaging,” *Signal, Image and Video Processing*, vol. 19, no. 5, Mar. 2025, doi: 10.1007/s11760-025-03959-2.
- [17] Y. Kim et al., “Early prediction of need for invasive mechanical ventilation in the neonatal intensive care unit using artificial intelligence and electronic health records: a clinical study,” *BMC pediatrics*, vol. 23, no. 1, Oct. 2023, doi: 10.1186/s12887-023-04350-1.
- [18] K. Yoneda, T. Seki, Y. Kawazoe, K. Ohe, and N. Takahashi, “Immediate postnatal prediction of death or bronchopulmonary dysplasia among very preterm and very low birth weight infants based on gradient boosting decision trees algorithm: A nationwide database study in Japan,” *PLOS ONE*, vol. 19, no. 3, p. e0300817, Mar. 2024, doi: 10.1371/journal.pone.0300817.
- [19] D. Rallis, M. Baltogianni, K. Kapetaniou, and V. Giapros, “Current Applications of Artificial Intelligence in the Neonatal Intensive Care Unit,” *BioMedInformatics*, vol. 4, no. 2, pp. 1225–1248, May 2024, doi: 10.3390/biomedinformatics4020067.

- [20] W. Jang et al., “Artificial Intelligence–Driven Respiratory Distress Syndrome Prediction for Very Low Birth Weight Infants: Korean Multicenter Prospective Cohort Study,” *JMIR. Journal of medical internet research/Journal of medical internet research*, vol. 25, pp. e47612-e47612, July 2023, doi: 10.2196/47612.