Listen to the Real Experts: Detecting Need of Caregiver Response in a NICU using Multimodal Monitoring Signals

Laura Cabrera-Quirós
lcabrera@itcr.ac.cr
Electronics Engineering, Costa Rican Institute of Technology
Cartago, Costa Rica

Gabriele Varisco
g.varisco@tue.nl
Applied Physics, Eindhoven University of Technology
Eindhoven, Netherlands
Clinical Physics, Máxima Medical Center
Veldhoven, Netherlands

Zhuozhao Zhan
Mathematics and Computer Science, Eindhoven University of Technology
Eindhoven, The Netherlands

Xi Long
Electrical Engineering, Eindhoven University of Technology
Eindhoven, Netherlands
Philips Research
Eindhoven, The Netherlands

Peter Andriessen
Applied Physics, Eindhoven University of Technology
Eindhoven, Netherlands
Pediatrics, Máxima Medical Center
Veldhoven, The Netherlands

Eduardus J.E. Cottaar
Applied Physics, Eindhoven University of Technology
Eindhoven, The Netherlands

Carola van Pul
Applied Physics, Eindhoven University of Technology
Eindhoven, Netherlands
Clinical Physics, Máxima Medical Center
Veldhoven, Netherlands

ABSTRACT

Vital signs are used in Neonatal Intensive Care Units (NICUs) to monitor the state of multiple patients at once. Alarms are triggered if a vital sign is below/above a predefined threshold. Numerous alarms sound each hour which could translate into an overload for the medical team, known as alarm fatigue. Yet many of these alarms do not require immediate clinical action of the caregivers.

In this paper we automatically detect moments that need an immediate response (i.e. interaction with the patient) of the medical team in NICUs by using caregiver response to the patient, which is based on the interpretation of vital signs and of nonverbal cues (e.g. movements) delivered by patients. The ultimate goal of such approach is to reduce the overload of alarms while maintaining the patient safety.

We use features extracted from the electrocardiogram (ECG) and pulse oxymetry (SpO_2) sensors of the patient, as most unplanned interactions between patient and caregivers are due to deteriorations. Since in our unit an alarm can only be paused or silenced manually at the bedside, we used this information as a prior for caregiver response. We also propose different labeling schemes for classification, each representative of a possible interaction scenario within the nature of our problem.

We accomplished a general detection of caregiver response with a mean AUC of 0.82. We also show that when trained only with stable and truly deteriorating (critical state) samples, the classifiers can better learn the difference between alarms that need no immediate response and those that do. In addition, we present an analysis of the posterior probabilities over time for different labeling schemes, and use it to speculate about the reasons behind some failure cases.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**; • **Applied computing** → *Health care information systems*; **Health informatics**.

KEYWORDS

NICU, critical alarms, patient-caregiver interaction, alarm fatigue, monitoring signals, ECG, machine learning

ACM Reference Format:
Laura Cabrera-Quirós, Gabriele Varisco, Zhuozhao Zhan, Xi Long, Peter Andriessen, Eduardus J.E. Cottaar, and Carola van Pul. 2021. Listen to the Real Experts: Detecting Need of Caregiver Response in a NICU using Multimodal Monitoring Signals. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461615.3485435

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8471-1/21/10.
https://doi.org/10.1145/3461615.3485435

Figure 1: Example of interaction between caregiver and patient as a response to a critical alarm.

1 INTRODUCTION

Monitoring alarms are used in Intensive Care Units (ICU) worldwide to alert the medical team of patient deterioration [20] and should be responded to by the medical team, to avoid physiological damage or even death [6, 10, 19, 22]. Monitored signals generally include heart rate (HR), blood pressure, and oxygen saturation in blood (SpO_2), among many others [6]. Critical alarms are generally accompanied by loud sounds to alert caregivers and are typically sent to a central station or sometimes to handhelds of caregivers [10, 23].

The response to critical alarms, shown with an example in Figure 1, is particularly important in Neonatal Intensive Care Units (NICUs), as this fragile population has deterioration moments that tend to escalate quite rapidly. This is because preterm infants have problems regulating their physiology, as their systems are not fully developed yet and they are not as well equipped as adults to self-restore during a deterioration [4, 23].

Nonetheless, not all alarms in a (N)ICU need to be responded to immediately if they are not yet considered clinically urgent [9, 13]. Other reasons behind the lack of response for some alarms include self-patient stabilization, a large number of false or irrelevant alarms (up to 70% of total number of alarms delivered to caregivers), or work overload in the medical team [7, 13, 17]. *Alarm fatigue* is the term used for the overload of monitor alarms, resulting in desensitization to alarms in caregivers (particularly in nurses) and could potentially lead to missing important ones [6, 14, 19]. The problem complicates further in decentralized environments, such as single-family rooms [23]. Alarm fatigue can become so problematic that it has been named a medical device technology hazard [14]. On the one hand clinically relevant alarms should be handled while, to reduce alarm fatigue, alarms not clinically relevant and requiring no action should be reduced [13, 18].

In this paper, we propose a method to detect which moments need to be responded to immediately by a caregiver, by using multimodal vital signs extracted from bedside patient monitors. Unlike current monitoring alarms which are based on values passing critical thresholds, we aim to distinguish between important deterioration events with clinical relevance and those that can be potentially ignored. To do so, we use the interaction of the medical team with the patient monitor as silencing and pausing monitor alarms can only be done manually next to the patient. During these caregiver-patient interactions, similar to other social phenomena, the medical team learns inherently to interpret nonverbal cues from neonatal patients (e.g. movements) and reacts accordingly. We hypothesize that this additional nonverbal information from the patients, along with the other parameters extracted from vital signs, allows the medical team to give a more accurate assessment of deterioration moments in their patients and what situations truly need a response.

This work aims therefore at improving current mechanisms for healthcare applications by revising the definition of alarms used in the NICU and substituting alarms based on thresholds with new alarms developed based on caregiver response to critical conditions recorded from their past experience. Furthermore, the solution of using multimodal vital signs together with caregiver responses as a response to the interpretation of patient signals and nonverbal cues from the patients is adopted in this work since it allowed to preserve privacy of both caregivers and patients in the NICU during the data collection and did not have an impact on working routines of caregivers. For instance, cameras were not used to record caregivers entering patient rooms as the use of these devices in intensive care setting could have resulted in a major privacy issue, a problem that we tackled with the alternative of using the interaction of the medical team with the monitor, and hence with the patient. This response can give important information on which alarms are useful and which ones can be filtered out. Reducing alarms can increase patient safety but also patient and family well-being. In addition, fewer alarms can lower stress levels for caregivers.

The main contributions of our work are:

- We automatically detect moments that need an immediate interaction of caregivers with patients in NICUs.
- We evaluate the response in moments where alarms were triggered by the bedside patient monitor, aiming to filter alarms that do not need an immediate response.
- We evaluate three different labeling schemes for training and testing of our models, based on the presence of monitor alarms and caregiver response. With each labeling scheme, we evaluate different scenarios of interaction that are inherent to the nature of our problem.

The rest of the paper is distributed as followed: Section 2 summarizes the related work. Section 3 details some important aspects in the day-to-day of our NICU which are relevant to better understand our problem and the steps taken to solve it. The dataset used in this work is described in section 4, while Section 5 presents the proposed method and our different labeling schemes. Section 6 shows the experimental cases created using our labeling schemes, and their result and discussion. Finally, our conclusions are presented in Section 7.

2 RELATED WORK

Some works discussed reducing alarm fatigue in NICUs by implementing delays between exceeding a threshold and the start of the alarm, or by using longer average periods for vital signs, to account for patient self-stabilization [16, 21]. Other heuristic approaches aimed towards wider limits for alarm thresholds [25] or customizing them to patient characteristics [18, 24]. However, these approaches

do not use interaction of the caregiver with the system to investigate clinical relevance of an alarm. Most works aiming to reduce false alarms and alarm fatigue in (N)ICUs focus on the vital signs. For example, Chen et. al. [3] used supervised machine learning to detect artifacts in online vital signs from multiple sensors. Similarly, Eerikäinen et. al. [8] focused on clinical and physiological signs of arrhythmia, in order to train classifiers capable of recognizing this specific deterioration event.

Previous work studied the assessment given by the medical team, especially nurses, to critical alarms and how this influences their reaction [9, 13, 18]. In particular, it was discussed that not all the alarms caused by vital signs exceeding thresholds triggered an immediate response by the medical team. Joshi et. al. [13] showed that nurses see all alarms as a good indicator, but they do not see the need to respond to all as only a fraction of alarms are considered urgent. More recently, Ergezen and Kol [9] provided an observational study to determine types of alarms and evaluate the response given by nurses to each type. Similarly, a rather low percentage of critical alarms (about 50%) was responded. Nevertheless, none of the above approaches used prior knowledge obtained from caregiver assessment of alarms to train detection or prediction models. The closest work to our own was presented by Ostojic et. al. [17]. Using HR and SpO_2 with four machine learning algorithms, they classified alarms responded to by the medical team as true or false alarms. However, the staff labeled only the moments that were responded and they did so in an offline manner. In contrast, in this work we assess all alarms generated by the patient monitor, and not only those responded to. Moreover, to the best of our knowledge, we are the first to use the assessment of caregivers in the onsite response to alarms to automatically detect moments that should be responded to, aiming for alarm reduction.

3 NICU MONITORING WORKFLOW

To better understand the nature of our problem, we describe the alarm handling workflow protocol in a NICU (see Figure 1). There are several types of monitor alarms in a NICU, classified given their severity and nature (e.g. sensor type or technical issue). Monitoring critical alarms due to a vital sign are triggered when the signal goes below or above a threshold predefined by the medical team. The value of these thresholds vary depending on the population (i.e. very low weight, low weight) and other specific characteristics of the patient [23].

A yellow alarm is triggered when a particular vital sign of the patient is outside of the stable range, but it is not yet life-threatening. These can deteriorate further or self-stabilize. Red alarms are triggered when a vital sign goes below/above critical levels and can threaten the life of the patient. These can also self-stabilize, but this scenario is less likely. As an example for our NICU, an SpO_2 value lower than 88% will trigger a yellow alarm while a value lower than 80% will produce a red alarm for desaturation.

All alarms produce alert sounds in the bedside monitors and are send to a central station, for medical information and assessment. As our NICU has a single-family room arrangement, red alarms are also sent to the wearable handle of the caregiver responsible for the patient. If no response is found after 45 seconds and the red alarm does not self-stabilize, this is send to the next caregiver available

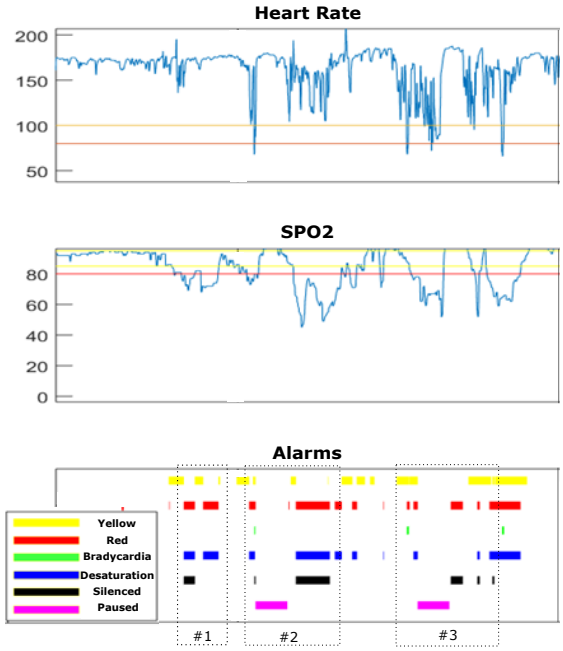

Figure 2: Alarm timeline for Heart Rate (HR) and Oxygen Saturation (SpO_2), with three different deterioration events.

following a well-defined hierarchy. When an immediate response is needed, the caregiver will go to the patient room to care for the patient and handle the alarm.

The first action in the room is to silence an alarm. Silencing means that alarms related to that sensor are silenced for 1 minute. For example, silencing a bradycardia alarm (low heart rate) would silence all alarms related to heart rate.

Another action could be to use a pause. When a pause is applied, all alarms are inactive for 3 minutes. Alarms can be paused while alarms are active and while there are no alarms. Pause is typically used when caregivers plan to take care of a patient (e.g. diaper change, feeding or kangaroo care[1]), knowing that these actions can disturb the sensors and lead to unnecessary alarms. The pause time can be stopped manually by the caregivers before the fading time, to prevent alarms being silenced or paused when the caregiver has left the bedside. Events like silence and pause therefore indicate that a caregiver is close to the patient, as this action cannot be performed in our unit without being at the bedside. We call these responded events.

Figure 2 shows an example of alarms for HR and SpO_2 for a deteriorating patient. The yellow and red alarms are subdivided into alarms for critically low heart rate (bradycardia) and SpO_2 (desaturation). Moments of silenced/paused alarms are also shown. The interaction of the caregiver with the patient was as follows. First, the patient deteriorated (low oxygen saturation levels) and that escalated to a red alarm which was silenced (event #1 in Figure 2). After this interaction, the infant reacted favorably and saturation

[1]Kangaroo care is a technique in which a caregiver puts the infant in his/her parents chest ensuring skin-to-skin contact, and it stimulates the infant's development [15]

Table 1: Patient demographics for our study.

Characteristic	Median	25th perc.	75th perc.
Gestational age (weeks)	29.4	28.71	30.71
Birth weight (g)	1260.0	1065.0	1417.5

Table 2: Used features per signal type. HRV: Heart Rate Variability. SpO_2: peripheral oxygen saturation. SDNN: Standard deviation of RR-intervals. RMSSD: square root of the mean of the sum of the squares of differences between adjacent RR-intervals.

Sensor	Feature group	Features
Electrocardiogram (ECG)	Heart Rate	mean, variance
	HRV	mean RR, SDNN, RMSSD
Pulse Oxymeter	SpO_2	mean, variance

increased again. The second deterioration event (#2) comes minutes later and consists of i) a desaturation followed by a bradycardia alarm which is considered more critical to the infant and ii) a pause of the alarms and several silencing episodes after the pause fading time is reached (3 minutes). This is considered a clinically relevant deterioration event. Finally, a third deterioration event occurred, similar to the event #2, after which the infant recovered and went back to stable ranges in the monitoring signals.

4 THE DATASET

Máxima Medical Center (MMC) is a general hospital in Veldhoven (The Netherlands) with a 22-bed (18 in use), level III tertiary NICU, and an admission rate of approximately 380 newborns per year. The NICU comprises 9 single rooms, 5 twin rooms and 1 triplet room. Prematurely born infants from 24 weeks and older can be treated. All patients are continuously monitored using the Philips IntelliVue MX 800 monitor and all signals and alarms are automatically stored in a private data warehouse (PIIC iX, Data Warehouse Connect, Philips Medical Systems, Andover, MA). The sample frequency of the ECG is 250Hz, and HR and SpO_2 are measured every second. The patient profile that determines thresholds for the red and yellow alarms is chosen based on the gestational age of the neonate.

The data for six days for a total of 66 preterm infants is retrospectively collected from the data warehouse. It comprises ECG, HR and SpO_2 signals, critical (red and yellow) alarms, and silenced and paused alarms. The patient demographics are presented in Table 1. The medical ethical committee of MMC provided a waiver for this study, in accordance with the Dutch law on medical research with humans (WMO).

About 5% of the time consists of yellow alarms, while 0.8% are red alarms. In addition, 1.6% of the time were either paused or silenced by the medical team. It is thus clear that we are dealing with a problem of imbalance classes. As such, the training of the classifiers and the metrics for its evaluation should be defined accordingly.

5 PROPOSED METHOD

Here we describe our approach to detect the need for caregiver response, based on clinically readily available data in the patient monitor. First, we describe the used sensors and the multimodal features derived from them. Then, we define a responded event and a deterioration event, and describe how they differ from raw critical monitoring alarms. Next, we present our different labeling schemes and the classifiers used in our experiments.

5.1 Multimodal Sensors and Features

We have two main sensors in this work: 1) the ECG to measure the electrical activity of the heart, and 2) the pulse oximeter for SpO_2. These sensors were chosen because the majority of critical alarms in a NICU comes from them [24]. We derived three different feature groups: 1) HR-based, 2) HRV-based and 3) SpO_2-based (see Table 2).

An RR-interval is the time elapsed between two successive peaks in the ECG, called R-peaks. The heart rate HR is determined from the RR-interval as the number of beats (peaks) per minute. This signal is pre-extracted in most modern patient monitors including the one in our dataset. The Philips IntelliVue MX 800 monitor adds a series of noise filtering and signal enhancing proprietary methods for a better estimation of the heart rate. We use the mean and the variance of the HR as our features for this modality as variance in the heart rate is indicative of incoming deterioration [11]. Heart Rate Variability (HRV) features are derived from this RR-interval signal [1, 2]. In the current work, we use the mean RR-interval value within a window (mean-RR), the standard deviation of the RR-intervals within a window (SDNN) and the square root of the mean of the sum of the squares of differences between adjacent RR-intervals (RMSSD) [2]. Finally, for the SpO_2 we simply calculate the mean and variance. These statistics are used since they have proven to be informative of deterioration [26] and since they were previously defined in other studies, allowing for an easier transfer of knowledge about results to caregivers.

5.2 Responded and Deterioration Events

Figure 2 shows three deterioration events. In our pipeline, first we find the alarms. Next, we cluster the silenced and paused alarms that were placed 3 minutes or less next to each other, as most interaction events between caregivers and the patient constitute a sequence of these alarms without separations longer than these, and it is also the longer fading time for pause/silencing in our system. Once the silenced and paused alarms are clustered into a single event, we investigate if critical red and yellow alarms are triggered within the event. Critical alarms that are closely placed together are also clustered in the same way as the pause/silence alarms.

We define a *responded event* as a cluster of closely placed silenced and/or paused alarms, which are our prior for caregiver response. If a responded event co-allocates or has closely placed red or/and yellow alarms, we call it a *deterioration event*. This way we can analyze the different types of such events (e.g. desaturation only, desaturation with bradycardia, arrhythmia). We add an additional 30 seconds at the beginning and end of each responded event, to account for caregiver response to the alarms (if any) while outside of the patient room and for stabilization after the event, respectively. Also, as a pause inactivates all alarms, we need to assess the seconds before a pause.

Figure 3 (top) presents the raw red and yellow alarms (in red), the alarms that were paused or silenced (black) and the responded events that resulted after the clustering scheme (blue). We can see

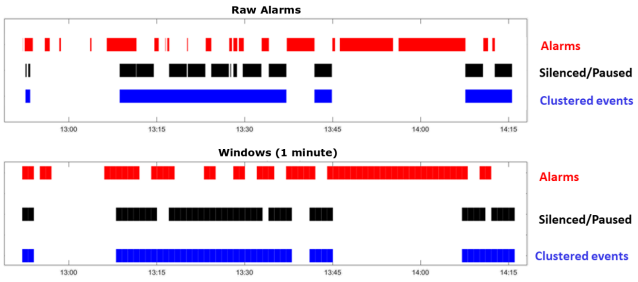

Figure 3: Clustering of paused/silenced alarms for the creation of responded events. If a responded event also has alarms within it, corresponds to a deterioration event. Top: using raw alarms. Bottom: using 1 minute windows.

in this figure that while there are 'holes' between paused/silenced alarms (black stream), these are filled after the clustering. In addition, it is clear from this figure that while critical alarms were triggered, these were not always responded.

The top section of Figure 3 presents the responded events derived from the raw alarms. However, the real start and end of any event is only known retrospectively. To adapt the creation of responded events for detection using machine learning, we present the same information in a sliding window manner (bottom of Figure 3, for windows of size T=1 minute without overlap). Thus, we also can have samples of consistently the same length. If there is at least one alarm (critical, silence or pause) within the window, that window is presented as having that type of alarm. Then, silenced/paused windows are clustered together to create the final responded events, and the critical alarm windows are evaluated for the deterioration events instead of the raw alarms. With this small change, the automatic detection of a caregiver response can be done for every window of T seconds.

5.3 Labeling Schemes and Detection of Caregiver Response

Each window of size T needs to be classified as responded or not. For all experiments, we use T equal to 1 minute. This choice was made to define windows long enough to include alarms and caregiver response (since time for response, in case this is present, is usually brief and less than 30 sec) while avoiding the inclusion of different consecutive alarms. For each window we extract the features summarized in Table 2 and label if alarms occur (Windows with Alarms as WA=1) and if an interaction with the monitor occurs (Window Responded as WR=1). These are the red stream (alarms) and the blue stream (events) in Figure 3. There are four possible combinations for WA and WR, resulting in one of four classes:

- **Class 1:** no alarms, no response (WA=0, WR=0)
- **Class 2:** alarms present, no response (WA=1, WR=0)
- **Class 3:** no alarms, response present (WA=0, WR=1).
- **Class 4:** alarms present, response present (WA=1, WR=1)

The first and last classes are clear cases where the patient is stable or during a deterioration event, respectively. On the contrary, the other two classes are rather ambiguous.

A response without alarms (class 3) generally represents routinely interactions such as caregiver handling that do not entail a deterioration. Thus, the features could potentially be closer to those in class 1, even while responded to. Similarly, alarms without response (class 2) could potentially be upcoming deterioration events that are not yet considered sufficiently relevant to respond, but for which the features can be closer to class 4. Class 2 could also represent momentary self-restoring instabilities, noisy signals or wrongly placed sensors that need no response at all.

Due to all the above plus the knowledge gathered at analyzing caregiver routines in a NICU, we create three different labeling schemes: 1) *complete* label, 2) *strict* label, and 3) *AlarmsOnly* label.

For the complete label, we only use the Window Responded (WR) stream as our ground truth which is the equivalent to detecting all responded events. Thus, the potentially ambiguous samples in classes 2 and 3 are forced to be part of class 1 and 4, respectively. With this label we aim to assess the global response of caregivers to the patient's needs, not only those due to deteriorations.

On the contrary, in the strict label we only consider classes 1 (stable) and 4 (deterioration events), for which we know for certain the state of the patient. For this labeling, which main use would be during training only (see Section 6), samples of classes 2 and 3 are discarded. The hypothesis behind this label is that it could allow the classifiers to learn the difference between a stable and a deteriorating infant, which can later translate to the necessity to respond or not to its aid during a deterioration event.

Finally, in the AlarmsOnly label we only consider classes 2 and 4, for which alarms were triggered within the window by the patient monitor. From a clinical point of view, the correct discrimination between classes 2 and 4 directly leads to a reduction of false alarms, thus reducing alarm fatigue. So this labeling scheme would allow us to properly assess such reduction.

We choose not to address our problem as a multiclass classification as we do not want to separate each class, but rather detect windows that should be responded to by the medical team. We do include the original four classes during our analysis to better explain possible reasons behind potential failure cases (Section 6).

5.4 Machine Learning Procedure

Using the three labeling schemes explained in Section 5.3, we create the six different experiments or cases summarized in Table 3.

The first case would be our general assessment of caregiver response, i.e. detection of responded events. Case 2 would only account for pre-selected samples, from which we can see if the classifiers can learn the difference between a stable patient and one with a clinically relevant deterioration. Hence, this case would also be our upper boundary. In addition, case 3 would focus on the particular cases where an alarm is triggered by the patient monitor, and the medical team should decide whether to respond to it.

Cases 4 to 6 represent meaningful training and testing combinations of the labeling schemes. First, in case 4 we evaluate the impact of learning the true separation between a deteriorating and a stable patient in the detection of all responded events. For case 5, we assess how this difference influences the necessity to respond to alarms triggered by the patient monitor. Finally, in case 6 we evaluate if training using all responded events can potentially be beneficial for the specific windows with an alarm triggered by the patient monitor. While the main goal of this work is to detect a

Table 3: Experimental cases. Each case represents a relevant combination during training/testing of our labelings.

	Labeling used	
CASE	Training	Testing
1	Complete	Complete
2	Strict	Strict
3	AlarmsOnly	AlarmsOnly
4	Strict	Complete
5	Strict	AlarmsOnly
6	Complete	AlarmsOnly

Table 4: Mean AUC (± standard deviation between folds) of the detection of caregiver response for the different cases.

Case	Logistic Regressor	Nearest Mean
1	0.821 ± 0.03	0.812 ± 0.03
2	0.935 ± 0.04	0.938 ± 0.03
3	0.783 ± 0.07	0.773 ± 0.07
4	0.791 ± 0.04	0.803 ± 0.04
5	0.772 ± 0.07	0.777 ± 0.07
6	0.777 ± 0.07	0.779 ± 0.07

window that should be responded (i.e. responded events), from a clinical point of view the most relevant cases are those for which the AlarmOnly label is used during testing (3, 5 and 6).

For each case, we applied a leave-5-subjects-out cross-validation using the same training and test data in all cases for a fair comparison. This means that all samples for 5 subjects (selected at random without replacement) are left out for testing in each fold, while the rest of samples are used for training.

We use a logistic regressor (L1 penalty) and a nearest mean classifier for all our experiments. We selected these to assess the problem with a linear separation and a clustering based approach, respectively. In addition, due to the simplicity that characterizes these two classifiers compared to other methods (e.g. neural networks), these classifiers allow to explain the importance of each feature, which is critical for the transfer of knowledge about results achieved with this work to caregivers. To consider the imbalance in our dataset during training, we weighted the contribution of each sample given its class presence during the optimization of the objective function [5].

For evaluation we used the Receiver Operating Characteristic curve (ROC) and the area under them (AUC). Unlike accuracy, these metrics are more suited to handle imbalanced data [5, 12]. In addition, we present confusion matrices for further evaluation.

6 RESULTS AND DISCUSSION

6.1 Classification Results

In Table 4 are summarized the mean (± standard deviation per fold) AUC for all the experimental cases. As expected, case 2 has the best performance (upper boundary) and shows that both classifiers can learn the difference between a stable and a deteriorating patient.

In addition, Figure 4 presents the mean ROC curves for all cases and the 2 classifiers. Table 5 presents the confusion matrices for one of the testing folds selected at random for the nearest mean classifier in case 1 (left) and case 4 (right). These are calculated for a threshold in the best operational point extracted from the ROC curves. The original classes as defined in Section 5.3 are also presented for further analysis. Note that the confusion matrices for cases 5 and 6 can also be derived from these matrices.

Results from Table 4 show that the highest AUC is found considering the strict labeling for both the training and testing dataset and the use of the logistic regressor (case 2). This result allows to associate a high true positive rate (0.90), main priority in a NICU due to complications associated to missed responses to critical conditions, with a very low false positive rate (0.20), indicating a very low incidence of unnecessary requested responses, a result significantly better compared to the high number of false or irrelevant alarms often reported in studies related to NICU environments [7, 13].

Case 1 shows that we can detect all responded events with a mean AUC of 0.82 for the logistic regressor, without much variance between the subject folds. The use of the strict label while training for the same goal (case 4) decreases slightly the AUC performance. Moreover, while the precision and specificity increases for the fold presented in Table 5 between cases 4 and 1, the sensitivity decreases as well. We speculate that by training with all possible events the classifiers can also discern other types of interactions, whereas with the strict label there is a focus on deterioration only.

This hypothesis is also supported by the original labels for each case, presented in Table 5. In the decomposed confusion matrix of case 1 in Table 5 (bottom left) is shown that from the 314 false negatives in the binary error only 35 are effectively from class 4 (a deterioration event). Most of the remaining 279 wrongly classified as no responded could potentially be routine interactions that do not entail a clinical deterioration (e.g. caregiver handling). For these cases there is a response (i.e. interaction with the patient) but the patient vital signs are stable during the interaction, confusing the classifier decision. The false negatives for class 3 further increase for case 4, for which a strict separation between stable and deteriorating samples was used in the training. While the true positives for class 4 also reduce between cases 1 and 4 by about 25%, this reduction was around 50% for class 3.

Unfortunately, there are no records in the dataset that explicitly show the type of interaction (e.g. deterioration, caregiver handling, parents) so we can not show the correct evaluation per class. Yet, this knowledge could help explaining several classification errors and gives us relevant insides for a future extension of this work.

When it comes to alarms triggered by the monitor, we can see in Table 4 and Figure 4 that there is barely any change between the ROC and AUC of cases 3, 5 and 6. However, analyzing the original labels of cases 1 and 4 in Table 5 (from which cases 5 and 6 can be derived) shows subtle differences for the classes 2 and 4.

For the 2584 samples of class 2 in this fold, which should be classified as not responded, the correctly classified samples go from 1688 for case 1 (sensitivity of 0.65 for the class) to 2271 for case 4 (sensitivity of 0.88). This behavior maintains in average across folds for the same classification threshold.

Moreover, when only considering class 2 and 4 in the confusion matrices, the results of caregiver response in this fold change from a sensitivity of 0.81, specificity of 0.65 and a precision of 0.15; to a sensitivity of 0.61, a specificity of 0.88 and a precision of 0.27. This evaluation is the equivalent of cases 6 and 5 (using the AlarmOnly during testing). This implies that training with the strict label reduced the number of alarms that needed no response but

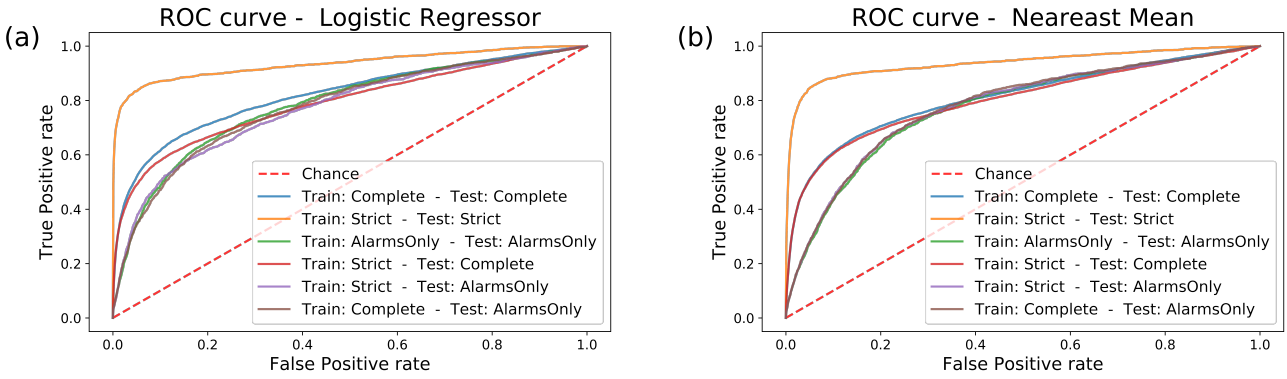

Figure 4: Receiver Operating Characteristic curves (ROC) for the detection of caregiver response for all experimental cases. (a) Logistic Regressor. (b) Nearest mean classifier.

Table 5: Confusion matrices for a random fold and cases 1 and 4 using the nearest mean classifier and a threshold equal to 0.65 (best operational point in ROC). Top: Confusion matrix for the case. Bottom: their respective true classes. Case 1: train with complete / test with complete. Case 4: train with strict/ test with complete.

Case 1

Complete Labeling	Estimated Labels		Totals
	No responded	Responded	
No responded	33218	2297	35515
Responded	314	371	685

True classes	Estimated Labels		Totals
	No responded	Responded	
Class 1	31530	1401	32931
Class 2	1688	896	2584
Class 3	279	218	497
Class 4	35	153	188

Case 4

Complete Labeling	Estimated Labels		Totals
	No responded	Responded	
No responded	34944	571	35515
Responded	453	232	685

True classes	Estimated Labels		Totals
	No responded	Responded	
Class 1	32673	258	32931
Class 2	2271	313	2584
Class 3	379	118	497
Class 4	74	114	188

were classified as responded, without drastically compromising the true positives. This unfortunately came with a reduction in the sensitivity that could also be explained by the classifier having a more strict concept of what is a deteriorating event.

6.2 Time Analysis of Probabilities

In addition to the global performance, we present the time analysis of the test data for 24 hours of one patient selected at random. Figure 5 presents the posterior probability of caregiver response when training with the complete label (top) and with the strict label (middle). The original class as presented in Section 5.3 is also shown (bottom). Each sample in this figure represent a window of one minute.

For case 1 (top) there are a several high probabilities around times for which alarms were triggered by the monitor but there was no response yet (class 2), as seen between hour 6 and 8 in Figure 5. In contrast, when training with the strict label, which only considers true deterioration events, the classifier is capable of reducing the probability for samples with only alarms while maintaining high probabilities for sections around a true deterioration (class 4).

This analysis helps showing that in a timely manner the classifier trained with the strict label can point to deterioration events in a general way, but has problems choosing the exact windows that overlap with the ground truth.

This mismatch could also explain the low precision values derived from Table 5, as these low performances are mostly due to the mismatches in windows around the deterioration events. As can be seen in Figure 5, the classifier for case 4 triggers a high probability around almost all deterioration events present in this day. However, the window-wise overlap between high probabilities and the ground truth is rather low for these segments.

We hypothesize that different window sizes or adapting the evaluation to consider this time dependency could potentially show fairer global results for precision. But we leave this for future efforts as it lies outside of the scope of this paper.

7 CONCLUSIONS AND FUTURE WORKS

In this paper we introduced our method to detect moments that need to be responded to by caregivers in a NICU, using multiple modalities extracted from signals in bedside patient monitors and

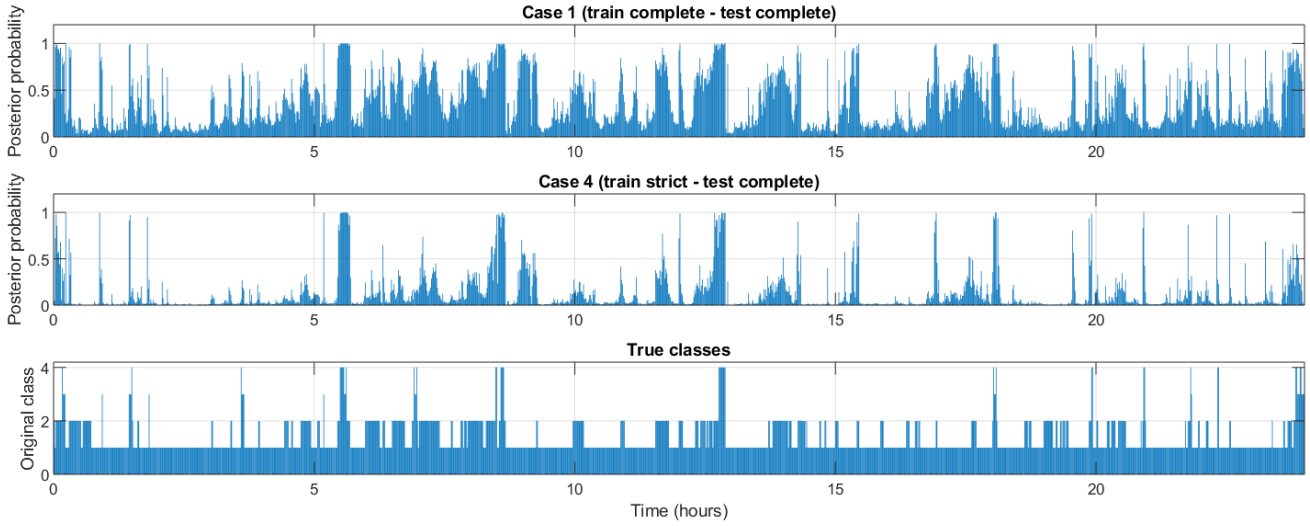

Figure 5: Time analysis of probabilities of the logistic regressor for one day of samples of a random test subject. (Top) Case 1: training with complete - test with complete. (Middle) Case 4: training with strict - testing with complete. (Bottom) Original classes.

caregiver response to alarms. This response comprises an inherently learned knowledge by the medical team of the nonverbal cues displayed by the patients, which combined with the information from monitoring signals better informs the medical team of the true state of the patients and their potential deteriorations.

We obtained a mean AUC of 0.82 for all responded events, a relevant result considering the high percentages of false alarms currently sounding in NICUs and causing alarm fatigue in caregivers. We used different labeling schemes for caregiver response, motivated in the type of interaction or response present in our data, both from a medical and a social point of view (e.g. kangaroo care). We showed that training to detect all responded events has a better general performance, but training only with samples of stable and truly deteriorating infants helps reducing the number of alarms triggered by the patient monitor that need no response, and are incorrectly classified as needed to responded.

We hypothesize that a better evaluation of types of interaction between caregiver and patients (e.g. further analyzing nonverbal cues displayed by the patients and their correlations with their health condition) can further help reducing our number of false positives. Thus, we could focus on deterioration events only. Additionally, incorporating time dependency in our algorithms could potentially improve the detection of responded events.

This study provides a first insight into the possibility of defining new meaningful alarms based on previously collected caregiver responses, as a prior for a more comprehensive understanding of the state of the patients that includes interpreting their nonverbal cues. Together with these results, future studies may shed a light on the possibility of revisiting alarms used in the NICU, by replacing threshold-based alarms with alarms defined based on previous caregiver responses, and possibly have an impact in reducing alarm fatigue in the NICU.

ACKNOWLEDGMENTS

The authors would like to thank patients' parents and medical staff of MMC for their support and enthusiasm while contributing with this work. This work was done within the framework of the Eindhoven MedTech Innovation Center (e/MTIC) which is a collaboration of the Eindhoven University of Technology, Philips Research and MMC. This work is a result of the ALARM project funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) grant number 15345.

REFERENCES

[1] U Rajendra Acharya, K Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit S Suri. 2006. Heart rate variability: a review. *Medical and biological engineering and computing* 44, 12 (2006), 1031–1051.

[2] A John Camm, Marek Malik, J Thomas Bigger, Günter Breithardt, Sergio Cerutti, RJ Cohen, Philippe Coumel, EL Fallen, HL Kennedy, RE Kleiger, et al. 1996. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation* 93, 5 (1996), 1043–1065.

[3] Lujie Chen, Artur Dubrawski, Donghan Wang, Madalina Fiterau, Mathieu Guillame-Bert, Eliezer Bose, Ata M Kaynar, David J Wallace, Jane Guttendorf, Gilles Clermont, et al. 2016. Using supervised machine learning to classify real alerts and artifact in online multi-signal vital sign monitoring data. *Critical care medicine* 44, 7 (2016), e456.

[4] Wei Chen, Sidarto Bambang Oetomo, and Loe Feijs. 2010. Neonatal monitoring: Current practice and future trends. In *Handbook of research on developments in e-health and telemedicine: technological and social perspectives*. IGI Global, 939–961.

[5] Davide Chicco. 2017. Ten quick tips for machine learning in computational biology. *BioData mining* 10, 1 (2017), 35.

[6] Maria Cvach. 2012. Monitor alarm fatigue: an integrative review. *Biomedical instrumentation & technology* 46, 4 (2012), 268–277.

[7] Barbara J Drew, Patricia Harris, Jessica K Zègre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. 2014. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PLoS One* 9, 10 (2014), e110274.

[8] Linda M Eerikäinen, Joaquin Vanschoren, Michael J Rooijakkers, Rik Vullings, and Ronald M Aarts. 2016. Reduction of false arrhythmia alarms using signal selection and machine learning. *Physiological measurement* 37, 8 (2016), 1204.

[9] Fatma Dursun Ergezen and Emine Kol. 2020. Nurses' responses to monitor alarms in an intensive care unit: An observational study. *Intensive and Critical Care Nursing* 59, 102845 (2020).

[10] Marjorie Funk, J Tobey Clark, Thomas J Bauld, Jennifer C Ott, and Paul Coss. 2014. Attitudes and practices related to clinical alarms. *American Journal of Critical Care* 23, 3 (2014), e9–e18.

[11] Alan H Gee, Riccardo Barbieri, David Paydarfar, and Premananda Indic. 2016. Predicting bradycardia in preterm infants using point process analysis of heart rate. *IEEE Transactions on Biomedical Engineering* 64, 9 (2016), 2300–2308.

[12] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[13] Rohan Joshi, Heidi van de Mortel, Loe Feijs, Peter Andriessen, and Carola van Pul. 2017. The heuristics of nurse responsiveness to critical patient monitor and ventilator alarms in a private room neonatal intensive care unit. *PloS one* 12, 10 (2017).

[14] James P Keller Jr. 2012. Clinical alarm hazards: a "top ten" health technology safety concern. *Journal of electrocardiology* 45, 6 (2012), 588–591.

[15] Deedee R Kommers, Rohan Joshi, Carola van Pul, Louis Atallah, Loe Feijs, Guid Oei, Sidarto Bambang Oetomo, and Peter Andriessen. 2017. Features of heart rate variability capture regulatory changes during kangaroo care in preterm infants. *The journal of pediatrics* 182 (2017), 92–98.

[16] Christina J McClure, S Young Jang, and Karen Fairchild. 2016. Alarms, oxygen saturations, and SpO 2 averaging time in the NICU. *Journal of neonatal-perinatal medicine* 9, 4 (2016), 357–362.

[17] Daniel Ostojic, Sabino Guglielmini, Virginie Moser, Jean-Claude Fauchère, Hans Ulrich Bucher, Dirk Bassler, Martin Wolf, Stefan Kleiser, Felix Scholkmann, and Tanja Karen. 2020. Reducing False Alarm Rates in Neonatal Intensive Care: A New Machine Learning Approach. In *Oxygen Transport to Tissue XLI*. Springer, 285–290.

[18] Halley Ruppel, Marjorie Funk, Robin Whittemore, Shu-Fen Wung, Christopher P Bonafide, and Holly Powell Kennedy. 2019. Critical care nurses' clinical reasoning about physiologic monitor alarm customisation: An interpretive descriptive study. *Journal of clinical nursing* 28, 15-16 (2019), 3033–3041.

[19] Sue Sendelbach and Marjorie Funk. 2013. Alarm fatigue: a patient safety concern. *AACN advanced critical care* 24, 4 (2013), 378–386.

[20] Sylvia Siebig, Silvia Kuhls, Michael Imhoff, Ursula Gather, Jürgen Schölmerich, and Christian E Wrede. 2010. Intensive care unit alarms—how many do we need? *Critical care medicine* 38, 2 (2010), 451–456.

[21] Ben J Stenson, William O Tarnow-Mordi, Brian A Darlow, John Simes, Edmund Juszczak, Lisa Askie, Malcolm Battin, Ursula Bowler, Roland Broadbent, Pamela Cairns, et al. 2013. Oxygen saturation and outcomes in preterm infants. *New England Journal of Medicine* 368, 22 (2013), 2094–2104.

[22] Carola van Pul, Rohan Joshi, Wouter Dijkman, Heidi van de Mortel, Jarno van den Bogaart, Thilo Mohns, and Peter Andriessen. 2015. *Alarm management in a single-patient room intensive care units*. IOS Press, 119–133.

[23] Carola van Pul, Heidi P M E van de Mortel, Jarno J L van den Bogaart, Thilo Mohns, and Peter Andriessen. 2015. Safe patient monitoring is challenging but still feasible in a neonatal intensive care unit with single family rooms. *Acta Paediatrica* 104, 6 (2015), e247–e254.

[24] Gabriele Varisco, Heidi van de Mortel, Laura Cabrera-Quiros, Louis Atallah, Dirk Hueske-Kraus, Xi Long, Eduardus JE Cottaar, Zhuozhao Zhan, Peter Andriessen, and Carola van Pul. 2020. Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU. *Acta Paediatrica* 100, 4 (2020), 1141–1150.

[25] James Welch. 2011. An evidence-based approach to reduce nuisance alarms and alarm fatigue. *Biomedical Instrumentation & Technology* 45, s1 (2011), 46–52.

[26] James R Williamson, Daniel W Bliss, David Browne, Premananda Indic, Elisabeth Bloch-Salisbury, and David Paydarfar. 2011. Using physiological signals to predict apnea in preterm infants. In *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. IEEE, IEEE, 1098–1102.