Machine learning applied to multi-sensor information to reduce false alarm rate in the ICU

Gal Hever[1] 0000-0003-1853-4310, Liel Cohen[1], Michael F O'Connor[2], Idit Matot[3], Boaz Lerner[1] and Yuval Bitan[1] 0000-0001-7053-7012

[1]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel
[2]Department of Anesthesia and Critical Care, The University of Chicago, Chicago, Illinois, USA
[3]Department of Anesthesia and Critical Care, Tel-Aviv Medical Center, Tel-Aviv, Israel

Corresponding author: Yuval Bitan, Department of Industrial Engineering & Management, Ben-Gurion University of the Negev, POB 653, Beer-Sheva, Israel. E-mail address: ybitan@bgu.ac.il ; Tel: +972-86472225

Conflict of interest: None.

Abstract

Purpose: Studies reveal that the false alarm rate (FAR) demonstrated by intensive care unit (ICU) vital signs monitors ranges from 0.72 to 0.99. We applied machine learning (ML) to ICU multi-sensor information, to imitate a medical specialist in diagnosing patient condition. We hypothesized that applying this data-driven approach to medical monitors will help reduce the FAR even when data are missing.

Methods: An expert-based rules algorithm identified and tagged in our database seven clinical alarm scenarios. We compared a random forest (RF) ML model trained on datasets with missing parameters (e.g., heart rate or blood pressure) in detecting ICU signals with the full expert-based rules (FER), our ground truth, and partial expert-based rules (PER), where missing parameters were removed from the rules.

Results: When all alarm scenarios were examined, RF and FER were identically perfect. However, in the absence of one to three parameters, RF maintained its values of the Youden index (0.94–0.97) and positive predictive value (PPV) (0.98–0.99), whereas PER lost its value (0.54–0.8 and 0.76–0.88, respectively). While the FAR for PER with missing parameters was 0.17–0.39, it was only 0.01–0.02 for RF. When scenarios were examined separately, RF showed clear superiority in almost all combinations of scenarios and numbers of missing parameters.

Conclusion: When sensor data are missing, specialist performance worsens with the number of missing parameters, whereas the RF model attains high accuracy and low FAR due to its ability to fuse information from available sensors, compensating for missing parameters.

Keywords: False Alarms, Intensive Care Unit, Machine Learning, Missing Data, Random Forest

Introduction

Monitoring patient clinical status is essential, particularly in intensive care units (ICUs). The monitoring systems in ICUs are designed to be highly sensitive. The standard for the implementation of these systems has been to maximize their sensitivity (e.g. the percentage of signals correctly identified) and accept the associated degradation of their specificity (e.g. the percentage of no-signals correctly identified). Thus, as thresholds for alarm triggering become more sensitive and less specific, more false alarms (FAs) are generated. This strategy for alarm initiation may predictably lead to an intolerable number of FAs [1].

Sendelbach & Funk [2] demonstrated that 0.72 to 0.99 of all alarms in critical care monitoring have no clinical relevance. Medical staff may be exposed to approximately 187 audible alarms [3] and 700 physiological monitor alarms [4] per day for each patient.

This huge number of FAs motivates healthcare professionals to turn down the volume of audible signals and ignore or even deactivate alarms [2]. Occurrence of multiple alarms at the same time makes it difficult to quickly identify an underlying condition [5] and inhibits clinician communication when this is most necessary [6]. Patient wellbeing is also compromised by the high levels of alarm noise in the room [7].

During the last decade, alarm hazards have been increasingly recognized as a major problem in the medical world. The Emergency Care Research Institute (ECRI) ranked the clinical alarm hazard as the top priority for the fourth year in a row in 2015 [8, 9, 10, 11], in the top 2 in 2016 [12], and in the top 3 in 2017 [13]. This reflects the severity of this problem and indicates a high priority for advances in the field.

The high false alarm rate (FAR) calls for the application of modern methods to identify and suppress FAR in real time. To encourage the development of algorithms to reduce FAR in the ICU, PhysioNet published the PhysioNet/Computing in Cardiology Challenge 2015 [14]. The organizers of this challenge focused on arrhythmia alarms for five life-threatening arrhythmia types and provided software resources and an open dataset that consisted of a collection of records of electrocardiogram, arterial blood pressure, and photoplethysmogram signals in which one of the five arrhythmia alarms occurred. Multiple approaches were analyzed to meet this challenge. Some investigations focused on machine-learning (ML) algorithms, some used signal-quality assessment or filtering methods, and others presented multivariate methods or a combination of the above. Eerikäinen et al. [15], who combined ML and signal-quality assessment methods, developed an algorithm that selects the two most reliable signals based on the F1-score [16], extracts features from both, and uses the F1-score as a classification evaluation measure.

Our research aimed at developing automated data-driven decision support tools using ML methodologies with the main goal to minimize the FAR of ICU vital signs monitors. This approach strives to imitate the way medical practitioners assess patient's condition by simultaneously comparing vital signs parameters from several sensors. In previous studies, this approach significantly reduced the FAR [17]. However, studies of alarms in the ICU have not dealt with missing data from sensors, which might frustrate the application of any set of expert-derived rules. Missing data are common and unavoidable in clinical care and might cripple the performance of any advanced rules that correlates information across sensors. Imhoff and Kuhls [18] presented the importance of criteria measuring robustness against missing values to enhance the clinical and technical quality of monitor alarms. Nevertheless, few studies of FARs in intensive care clinical data considered missing values. Vesin et al. [19] found that out of 44 published clinical studies, 16 made no mention of missing data. Moreover, less than 5% acknowledged the importance of missing data and the need to account for it in real time.

Thus, we designed this study to analyze in real-time missing sensor data to minimize false alarm rate. We hypothesized that applying an ML data-driven approach toward developing a clinical alarm model would reduce FAR reported in ICUs in situations of missing parameters.

Methods

Model Overview

Bitan & O'Connor [17] demonstrated an approach for suppressing FARs using expert-based rules. This approach correlates information across sensors and improves the performance of alarms by replicating specialist reasoning. They found that FARs were reduced when comparing parameters across sensors. They also evaluated more sophisticated rules of alarm scenarios by correlating information across sensors. However, these expert-based rules are heavily dependent on physiological data collected by separate sensors. Missing data from these sensors degrades the performance of the algorithms. Furthermore, the inability of an algorithm to deal with missing sensor data causes an increase in both FARs and false negatives, which decreases the algorithm's effectiveness.

Therefore, in the present study, we addressed the missing data phenomena in critical care using ML, which can consolidate information from available sensors to compensate for missing sensors. We trained random forest (RF) models based on clinical parameters to mimic decisions based on the full expert-based rules (FER) (Table 1). In its training, the RF samples the parameter-based clinical database randomly *n* times, and for each derived data set (sample), it learns a tree classifier that at each split in its learning process randomly selects a subset of size *k* of the parameters. During the test, prediction for a new patient is made according to the majority vote of the *n* tree classifiers. The values of *n* and *k* that maximize classification performance are selected during the training using a data set that is independent of the training and test sets (see *Model Implementation*). The classification made by the FER, which is equivalent to that of the medical specialist, was used as a ground truth in model training and performance evaluation (we do not know if the clinicians caring for these patients made the same diagnosis as specified by the FER, but it would have been reasonable for them to do so). We then trained RF models based on partial sets of parameters, i.e., without one, two, or three of the parameters. In these partial settings, the FER classification was also used as a ground truth. The classification test results of the RF models trained using the partial sets of parameters were compared with those of the compatible partial expert-based rules (PER) to check which of the two – the medical specialist or the ML-based model, both relying on missing data – was more accurate.

The partial models were stored in a "bank" of alarm classifiers that can be used for detecting alarms in ICU monitoring systems, even when the measurement of several parameters has failed. For example, if the patient’s arterial blood pressure systolic (ARTBPS) parameter is missing in real time, and the medical staff must make a decision without it, a model trained on data that does not include the parameter could be extracted from the bank and applied at the bedside. The classification of the extracted model will be based on correlative parameters with ARTBPS. That is, using ML, parameters that track with ARTBPS can be identified and used in its place when it is not available; in this situation, the alarm might have a performance that is inferior than it would have using ARTBPS, but can be surprisingly close. During and as part of training, the RF selects an optimal combination of parameters for prediction that even if one or more of them are missing, the classifier remains accurate overcoming the missingness of that parameter thanks to the others in the combination.

This pre-trained model, which does not require ARTBPS as an input parameter, may be used in the monitoring system to assist the medical staff in assessing a patient's condition. Also Eerikäinen et al. [15] used a bank of classifiers, each for a different clinical alarm, but they did not tackle parameter missingness by extending the classifier bank for each combination of missing parameters, as we did, but used the F1-score [16] to select those parameters that were most reliable.

We used two main performance measures to optimize and evaluate the RF models: positive predictive value (PPV) and Youden's index. PPV (also called precision) is the percentage of samples correctly specified as positive [20]:

$$\text{PPV} = \frac{\text{True positives}}{\text{True positives}+\text{False positives}}$$

The PPV measure of a medical monitor is the percentage of justifiable alarms. This term is useful since it answers the question: “When the alarm sounds, how often is it correct?” [21].

Youden's index (*J*) combines sensitivity and specificity into a single measure. Searching for the model that maximizes this index allows us to find the one that maximizes the contribution to both sensitivity and specificity [22]:

$$J = max_c\{Sensitivity(c) + Specificity(c) - 1\}$$

where *c* is the discrimination threshold (or cut-off point) of the model’s output. Note that our approach not only evaluates the classifier using performance measures relevant to critical care as in other studies (see, e.g., Eerikäinen et al. [15], whose evaluation used sensitivity, specificity, and “weighted” accuracy), but also augments it during training (optimization) to maximize these measures in the test. Specifically, we selected the RF that maximizes the Youden’s index, which guarantees maximization of the sum of sensitivity and specificity, as well as when classifying never-before-seen examples. For the completeness of presentation, in the Results section, besides reporting on Youden’s index, we also report classifier performance using other most common measures (PPV, sensitivity, specificity, and FAR).

Data Source

This is a retrospective study that was approved by the Tel-Aviv Medical Center (TAMC) ethical board (0673-15-TLV). The data for this study were retrieved from samples of anonymous patients who were admitted to the TAMC ICU between 2008 and 2014. Data were collected using Metavision software, which tracked information from a patient’s bedside monitor in order to calculate and store the average per minute value of the patient’s vital signs. The monitors at the TAMC ICU belong to the DATEX series of F-CUB 08 models and can have a high sampling frequency (hundreds of samples per second).

The database contained 681,265,089 data points obtained from 7,688 patients. Different patients were monitored for different sets of physiological parameters. Some parameters were measured for all patients, while others were measured only for a few. We elaborate on the choice of parameters for our model in the following section.

Model Implementation

This research applied expert-based rules to tag clinical alarms and diagnosed patient medical state using a data-driven model trained to map multiple signals across monitor sensors onto the tagged alarms. According to the suggested FER approach, an alarm is triggered when one or more parameters cross their threshold values at a predetermined time. Thus, each sample in the dataset was tagged as "Alarm" or "No alarm" according to a set of classification rules defined by a specialist doctor based on the measured values of five parameters: heart rate (HR), ARTBPS, arterial blood pressure mean (ARTBPM), central venous pressure (CVP), and pulmonary arterial diastolic pressure (PAPD). We considered alarms triggered in seven clinical scenarios: Bradycardia, Bradycardia hypotension, Hypovolemia, Tachycardia, Tachycardia hypotension, Obstructive shock, and Left ventricular (LV) shock. Table 1 presents the parameters and their thresholds, establishing together the FER for clinical alarms in a time window of one minute for each of the seven scenarios. Besides being the gold standard in the experiments, these FERs produced the labels (ground truth) for training our ML algorithm.

Clinical alarm scenario	**Rules**
Hypovolemia	ARTBPM < 50 mm Hg and CVP < 5 mm Hg and CVP > -10 mm Hg
Obstructive shock	ARTBPS < 78 mm Hg and CVP > 16 mm Hg and CVP < 35 mm Hg and PAPD > 16 mm Hg and PAPD < 60 mm Hg
LV shock	ARTBPS <78 mm Hg and CVP < 16mm Hg and PAPD > 16 mm Hg and PAPD < 60 mm Hg
Bradycardia	HR < 45 bpm
Bradycardia hypotension	HR < 45 bpm and ARTBPS < 78 mm Hg
Tachycardia hypotension	HR > 120 bpm and ARTBPS < 78 mm Hg
Tachycardia	HR > 120 bpm

Table 1—The full expert-based rules that composed the ground truth for this study. Note: We used the heart rate from the EKG as HR, as other heart rate signals were not available in our dataset.

FERs employ physiological parameters that are sampled by different sensors; however, in an unsynchronized manner, and often with varied sensor availability (e.g., a sensor detaches or fails), which leads to missing data and, consequently, to high FAR. To deal with missing data (parameters), a popular ML boosting and ensemble-learning algorithm, RF, was used. To train the RF, we extracted a dataset from the clinical database that contained 20,045 samples without missing values. Each such sample portrays simultaneous measurement of 12 physiological parameters for a single patient at a specific minute. These extracted parameters were: HR, ARTBPS, ARTBPM, CVP, PAPD, RR (respiratory rate) total, RR mandatory, saturation peripheral oxygen (Spo2), fraction of inspired oxygen (Fio2), ST1 (ST segment; the part of an electrocardiogram between the QRS complex and the T wave), ST2, and ST3. Some of the parameters were chosen because of their medical importance (as part of the FERs presented in Table 1) and others because of their high prevalence in the ICU recordings. Table 2 lists them with their frequency in the TAMC database and their percentage of missingness. The percentage of missingness was calculated by the ratio between the number of empty samples for patients for whom the parameter was sampled at least once and the total number of samples for those patients.

	Parameter name	% of patients	% of missing data
Always monitored parameters	HR	99.83	1.70
	Spo2	99.66	9.17
	ST1	97.06	8.69
	RR Total	94.66	56.58
Frequently monitored/charted parameters	ST2	88.85	21.90
	ST3	88.85	21.90
	Fio2	88.18	60.31
	ARTBPS	80.06	17.77
	ARTBPM	80.39	17.80
	RR Mandatory	79.76	56.12
Less frequently monitored parameters	CVP	24.88	64.98
	PAPD	3.91	85.06

Table 2—Parameter frequency and missingness in the TAMC database.

In order to investigate the situations of all the clinical alarm scenarios combined and each one individually, we conducted two tests. For the first test, we extracted a dataset that contained 1,500 events for each of the alarm scenarios. This number was chosen according to the number of events in the smallest out of the seven clinical scenarios datasets. In the second test, samples of the seven alarm scenarios composed seven separate datasets. The description of the following steps is relevant for both tests.

To minimize bias and variance in learning a classifier, the datasets were divided according to the ML methodology of cross validation (CV) [23] into three approximately equal-sized subsets, which were alternatively used for training, validating, and testing the RF models.

During the training phase, RF models containing between 40 and 200 trees in each forest and 1 to 7 parameters in a tree were evaluated using Youden's index (where the alarm and no alarm classes were sampled to balance the natural imbalance of the dataset in favor of the no alarm class). The output of a model provided the estimated probability for alarm triggering for each of the 12-dimensional samples. A threshold value between 0 and 1 was set such that a probability greater than this threshold indicated an "Alarm" event and below (or equal) it, a "No alarm" event. The RF model that achieved the highest Youden’s index value on the validation set among RF models trained according to all possible thresholds was selected for the test. To evaluate the test performance (estimating the accuracy of alarm classification in real time), an average was calculated over the results for the three test subsets derived in the CV experiment, presented in the Results section. Using Youden’s index while training and validating the candidate RF models, and not only in evaluating (testing) the models, guaranteed that the selected RF, having specific numbers of trees and parameters, was augmented from the outset toward maximizing the sum of sensitivity and specificity.

As previously mentioned, FER was used as the ground truth for the training process, both for partial and full RF models (i.e., models that use a partial or full set of parameters). The test results of the RF models were compared with the classification results of the FER and PER, where the latter is FER excluding rules that include missing parameters. This allowed us to compare the decisions made by a medical expert (FER/PER) to those by an automated ML model that was data-driven, mimicking the expert decision making.

The tests were implemented using the R package *randomForest* [24] using Rstudio software version 3.4.1. The source code is available on the GitHub repository: https://github.com/galhev/Machine-learning-applied-to-multi-sensor-information.

Results

Table 3 presents the average performance measures in the first test (all the clinical alarm scenarios together) for models with up to three missing parameters. The table shows that while none of the parameters is missing, the difference between RF and FER is negligible. However, when some parameters are missing, RF is clearly superior to PER with respect to all performance measures. For example, the RF average FAR is considerably lower for one (0.01), two (0.02), and three (0.02) missing parameters than PER (0.17, 0.29, and 0.39, respectively). The full results (Appendix A) show that of the 63 combinations of tests for one (7), two (21), and three (35) missing parameters, in four (6.34%) of the cases, PER achieved better scores than RF, in six (9.5%) of the cases, RF and PER achieved the same score, and in 53 (84%) of the cases, RF was superior to PER.

# of missing parameters	Model	Youden's index	PPV	Specificity	Sensitivity	FAR
None	RF	0.99	0.99	0.99	0.99	0.01
	FER	1	1	1	1	0
One	RF	0.97	0.99	0.99	0.98	0.01
	PER	0.8	0.88	0.83	0.97	0.17
Two	RF	0.96	0.98	0.98	0.97	0.02
	PER	0.66	0.81	0.71	0.95	0.29
Three	RF	0.94	0.98	0.98	0.95	0.02
	PER	0.54	0.76	0.61	0.92	0.39

Table 1—Average model performance with and without missing parameters. Note that PER for no missing data is the FER, which is our reference/benchmark result. Green and orange cells indicate better and worse results, respectively.

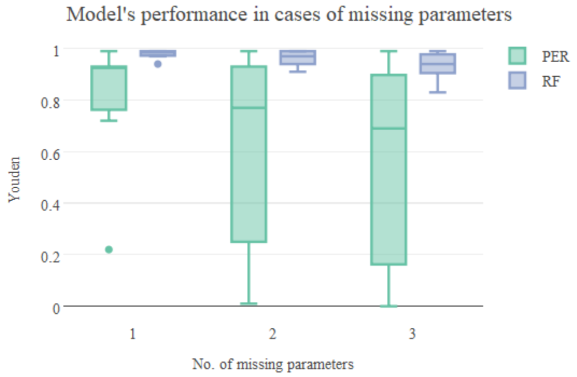

Figure 1: Model's Youden indices for missing parameters.

Figure 1 presents the Youden's index for the full results of Appendix A in a boxplot. The graph shows that PER performance becomes dramatically poorer as more parameters are missing from the data, in contrast with RF that maintains high sensitivity and specificity almost regardless of the missingness.

The second test used seven different datasets, each containing samples for a different alarm scenario. We repeated the procedure applied above to each scenario separately. Table 4 presents the average performance measures for each model (FER, PER, and RF) in each clinical alarm scenario for a different number of missing parameters. Note that the reported missing parameters in Table 4 for a clinical alarm scenario are those that established this scenario, i.e., making the FER (Table 1). For example, Hypovolemia was established by ARTBPM and CVP (Table 1); therefore, the reported missing parameters for Hypovolemia are ARTBPM, CVP, and their combination.

Scenario	# of alarm events	Missing parameters	Model	Youden's index	PPV	Specificity	Sensitivity	FAR
Bradycardia hypotension	653 out of 1,354	No	RF	0.98	1	1	0.98	0
			FER	1	1	1	1	0
		HR	RF	0.97	0.98	0.98	0.97	0.02
			PER	0.7	0.92	0.88	0.82	0.12
		ARTBPS	RF	0.98	1	1	0.98	0
			PER	0.39	0.58	0.39	1	0.61
		HR + ARTBPS	RF	0.96	0.87	0.78	0.99	0.22
			PER	-	-	-	-	-
Bradycardia	676 out of 1,422	No	RF	0.99	1	1	0.99	0
			FER	1	1	1	1	0
		HR	RF	0.45	0.7	0.59	0.68	0.41
			PER	-	-	-	-	-
Tachycardia hypotension	1,678 out of 3,412	No	RF	1	1	1	1	0
			FER	1	1	1	1	0
		HR	RF	0.99	0.99	0.99	1	0.01
			PER	0.78	0.85	0.78	1	0.22
		ARTBPS	RF	0.99	1	1	0.98	0
			PER	0.27	0.62	0.27	1	0.73
		HR + ARTBPS	RF	0.78	0.72	0.65	0.95	0.35
			PER	-	-	-	-	-
Tachycardia	1,559 out of 3,067	No	RF	0.99	1	1	0.99	0
			FER	1	1	1	1	0
		HR	RF	0.43	0.71	0.41	0.64	0.59
			PER	-	-	-	-	-
Hypovolemia	1,104 out of 2,363	No	RF	1	1	1	1	0
			FER	1	1	1	1	0
		ARTBPM	RF	0.94	0.97	0.97	0.97	0.03
			PER	0.89	1	1	0.89	0
		CVP	RF	1	1	1	0.99	0
			PER	0.99	1	1	0.99	0
		ARTBPM + CVP	RF	0.8	0.7	0.55	0.95	0.45
			PER	-	-	-	-	-

Scenario	# of alarm events	Missing parameters	Model	Youden's index	PPV	Specificity	Sensitivity	FAR
Obstructive shock	546 out of 1,214	No	RF	0.99	1	1	0.99	0
			FER	1	1	1	1	0
		ARTBPS	RF	0.89	0.91	0.83	0.9	0.17
			PER	0.25	0.59	0.25	1	0.75
		CVP	RF	0.99	0.99	0.99	0.99	0.01
			PER	0.97	0.99	0.98	0.99	0.02
		PAPD	RF	0.99	0.99	0.98	0.99	0.02
			PER	0.86	0.93	0.89	0.97	0.11
		ARTBPS + CVP	RF	0.76	0.65	0.46	0.96	0.54
			PER	0.21	0.52	0.2	0.95	0.8
		ARTBPS + PAPD	RF	0.91	0.86	0.84	0.95	0.16
			PER	0.04	0.54	0.04	1	0.96
		CVP + PAPD	RF	0.98	0.87	0.78	0.9	0.22
			PER	0.86	0.9	0.87	0.97	0.13
		ARTBPS + CVP + PAPD	RF	0.64	0.71	0.61	0.64	0.39
			PER	-	-	-	-	-
LV shock	1,786 out of 3,657	No	RF	1	1	1	1	0
			FER	1	1	1	1	0
		ARTBPS	RF	0.8	0.8	0.83	0.91	0.17
			PER	0.25	0.8	0.28	0.97	0.72
		CVP	RF	0.99	0.99	0.97	1	0.03
			PER	0. 98	0.99	0.98	1	0.02
		PAPD	RF	0.97	0.98	0.95	1	0.05
			PER	0.85	0.89	0.9	0.95	0.1
		ARTBPS + CVP	RF	0.71	0.82	0.77	0.82	0.23
			PER	0.23	0.68	0.48	0.75	0.52
		ARTBPS + PAPD	RF	0.86	0.91	0.91	0.87	0.09
			PER	0.02	0.57	0.02	1	0.98
		CVP + PAPD	RF	1	1	1	1	0
			PER	0.59	0.78	0.77	0.82	0.23
		ARTBPS + CVP + PAPD	RF	0.68	0.85	0.55	0.9	0.45
			PER	-	-	-	-	-

Table 2—Average performance measures of RF and FER/PER for different clinical alarm scenarios and missing parameter settings. Green and orange cells indicate better and worse results, respectively, for each of the missing parameters in each clinical alarm scenario. An empty cell indicates the inability to exercise the PER model due to the missing parameter(s) this model is based on.

Missing results for PER (indicated by “-” in Table 4) reflect situations of inability to exercise the PER model due to missing parameters on which the rules are based. In cases where all parameters composing a specific scenario were missing, PER could not function. For example, if the HR parameter was missing, it was impossible to exercise the single rule, HR<45, which is needed to detect Bradycardia. In such situations, the RF, integrating and fusing information from the available parameters compensating for those missing, has a significant advantage over the PER, which is not applicable. For example, looking again at the Hypovolemia scenario, when the parameters that composed its rules (ARTBPM and CVP) were missing, PER was not applicable (“-” in Table 4), and thus was useless. However, RF still succeeded in producing a good Youden's index and PPV measure of 0.8 and 0.7, respectively, using the remaining parameters, which provided complementary information about the patient status albeit the absence of ARTBPM and CVP. Similarly, while PER was not applicable in detecting Bradycardia hypotension alarms in the absence of HR and ARTBPS, RF succeeded in this task with relatively good performances (Youden’s index of 0.96 and PPV of 0.87), thanks to extracting information from the remaining (complementary) parameters. More about this important advantage of RF follows next.

According to Table 4, the reduction in FAR due to the use of RF compared to PER with one or two parameters missing for Obstructive shock is between 0.01 and 0.8, except in the case where CVP and PAPD were missing, and PER achieved a better FAR (0.13) than RF (0.22). When all three parameters used to compose the rules of Obstructive shock were missing, PER was not applicable, whereas RF achieved a FAR of 0.39. For Bradycardia and Tachycardia when HR was missing, PER was again not applicable, and RF achieved FARs of 0.41 and 0.59, respectively. The reduction in the FAR due to the use of RF compared to PER for Bradycardia hypotension and Tachycardia hypotension with one parameter missing was 0.10 or 0.61 and 0.21 or 0.73, respectively. When the two parameters establishing the scenario were missing, the RF FARs were 0.22 and 0.35, respectively, whereas PER was completely not applicable. The reduction in the FAR for the LV shock clinical alarm scenario by RF was the highest of all alarm scenarios tested (from 0.05 up to 0.89 when ARTBPS and PAPD were missing) except for the case where CVP was missing, where PER (0.02) and RF (0.03) achieved comparable results. In the case where all the parameters used to compose the rules for LV shock were missing, RF achieved a FAR of 0.45, but PER was not applicable. The Hypovolemia FAR was equal for RF and PER except in the case in which ARTBPM was missing, where PER achieved a better score (0) than RF (0.03), and in which all the parameters were missing, where RF achieved a FAR of 0.45, but PER was not applicable. In summary, in both the first (Table 3) and second (Table 4) tests, RF presented a significant advantage over PER and a negligible difference with FER (the ground truth).

Finally, we identified the most important parameters for classifying each clinical alarm scenario by RF using the Gini index [25]. This index measures the impurity of a dataset by considering a binary split for each parameter and selecting the parameter providing the highest reduction in impurity for the next classification step (note that “highest reduction in impurity” is equivalent to “making a classification with the lowest number of errors”). Computing the Gini index for each parameter in the set of all parameters in the first RF classification step, the one with the highest reduction in impurity is considered the most important for classification and is selected first. Then, in the second classification step, the same procedure is applied to the set without the one already selected in order to identify the second most important parameter and so forth. This selection procedure creates an order of parameters ranked according to their importance in classifying a clinical alarm scenario. Table 5 presents the four parameters that achieved the highest ranking by RF according to this procedure. For example, when none of the parameters used to compose the rules of Hypovolemia is missing (i.e., the full set of parameters), ARTBPM achieves the highest reduction in impurity for classifying the dataset and, thus, is considered the most important parameter to Hypovolemia. Then when ARTBPM is missing (column "One"), the next parameter that achieves the highest reduction in impurity (i.e., the second most important parameter for Hypovolemia) is CVP and so forth.

According to Table 5, the first most important parameters to classify most of the alarm scenarios (where none of the parameters was missing) are HR and ARTBPS. By comparing Tables 5 and 1, we see that RF always correctly detected one of the parameters establishing the rules of a clinical alarm scenario (FER) as the most important and, in most cases, also the second most important parameter according to these rules (except for LV shock). Parameters that were not in the FER (Table 1) were discovered later by RF, only when the parameters in the FER became missing (Table 5), showing the ability of the RF to compensate for parameter missingness and to provide accurate classification, albeit having missing parameters. For example, for Tachycardia, when HR was missing, the next parameter that compensated for this missingness was ARTBPS (which is not part of the FER for Tachycardia; Table 1), and the next important parameter when also ARTBPS was missing was ST1 (also not part of the FER). This compensation makes sense in light of the fact that all three parameters are related to contraction of the heart.

# of missing parameters / Alarm scenario	None	One	Two	Three
Hypovolemia	ARTBPM	CVP	ARTBPS	PAPD
Obstructive shock	ARTBPS	CVP	ARTBPM	PAPD
LV shock	ARTBPS	ARTBPM	PAPD	CVP
Bradycardia	HR	ARTBPS	ST1	RR mandatory
Bradycardia hypotension	ARTBPS	HR	ARTBPM	ST1
Tachycardia	HR	ARTBPS	ST1	RR mandatory
Tachycardia hypotension	HR	ARTBPS	ARTBPM	ST1

Table 3—The parameter achieving the highest ranking according to the Gini index applied to the RF classification results for each scenario and a different number of missing parameters.

Discussion

Vital signs monitors are an important part of patient care. They allow medical professionals to assess and keep track of a patient's condition and progress. Medical monitors in use today in the ICU track a patient's status by reporting data from each sensor separately. Unfortunately, the FAR of such monitors is staggering, to the point where they are nearly useless (hence the ECRI priority).

Previous studies intended to reduce FAR utilized an approach driven by expert decision assessing the patient's state using multiple sensors instead of one. This approach produced more sophisticated rules of alarm scenarios and held enormous potential to reduce the FAR. However, the performance of this approach degrades quickly in the face of missing sensor data, which are common in the ICU. The more sensor data that are missing, the less useful this approach becomes.

In this study, we compensated for missing data by applying a ML method to multi-sensor information. We used the RF algorithm to replace a set of expert-based rules for identifying clinical alarms by utilizing complex relations in the multi-sensor information to establish and train a pre-stored "bank" of classifiers, one for each missing sensor/parameter. Such pre-trained models can be used in ICU monitoring, where a missing sensor/parameter will trigger application of the corresponding classifier already trained for this missingness to assist the medical staff in diagnosing a patient's condition. Based on several performance measures, our results demonstrate that the utility of RF for overcoming missing sensor data is appreciable, typically performing better than any single sensor yet studied. Hence, we suggest a simple and practical solution for suppressing FAR significantly in situations of missing sensor data and reducing alarm fatigue in ICUs.

Sample tagging in our study was based on FER, but due to lack of resources, this was not validated by a doctor and was not tested in a patient setting. While it is unlikely that because of this we missed a large number of clinical events, our results would benefit from clinical validation.

Conclusions

The technology for building knowledge-based systems by inductive inference from examples has been demonstrated successfully in several practical applications. In this work, we present an approach for FAR reduction that improves patient safety by creating a quieter and more reliable ICU environment. This approach also creates a more suitable work environment for healthcare professionals and minimizes the negative effect of alarm fatigue.

Our data-driven machine learning approach defends alarm performance, even when some sensor (parameter) data are missing. When no data are missing, the system’s performance is comparable to a human medical expert. In cases of data loss of one or more physiological parameters, the machine learning derived alternatives degrade much less quickly than partial expert rules do. Thus, this might be considered as an alternative to the existing clinical alarm system.

To summarize, our approach can be used to improve patient monitoring in ICUs and increase alarm specificity using a data-driven multivariate method, which synchronously fuses data sources to make informed decisions. As demonstrated, this approach can work well on unseen data without the need of readjustments. Training the models in advance based on a specialist's rules enables the extraction in real time of the appropriate model from a bank of classifiers according to the current situation, which may also include missing sensors, without increasing the FAR. Note also that our approach allows an operating point to be selected for each individual patient by changing the threshold set on the classifier to balance the desired sensitivity and specificity for that patient, and to suit specific qualities such as age, gender, and known medical conditions. Finally, this approach can also be applied to other critical care units and extended to other medical devices, further aiding clinicians.

References

[1] Drew BJ, Califf RM, Funk M, Kaufman ES, Krucoff MW, Laks MM et al (2004) Practice standards for electrocardiographic monitoring in hospital settings. Circ 110(17):2721-2746

[2] Sendelbach S, Funk M (2013) Alarm fatigue: a patient safety concern. Adv Crit Care, 24(4):378-386

[3] Drew BJ, Harris P, Schindler D, Salas-Boni R, Bai Y, Tinoco A et al (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):1-23

[4] Cvach M (2012) Monitor alarm fatigue: an integrative review. Biomed Instrum Technol 46(4):268-277

[5] Sorkin RD (1988) FORUM: Why are people turning off our alarms? J Acoust Soc Am 84(3):1107-1108

[6] Edworthy J (1994) The design and implementation of non-verbal auditory warnings. Appl Ergon 25(4):202-210

[7] Xie H, Kang, J, Mills, GH (2009) Clinical review: the impact of noise on patients' sleep and the effectiveness of noise reduction strategies in intensive care units. Crit Care 13(2):208

[8] ECRI Institute (2011) Top 10 heath technology hazards for 2012. Health Devices 40(11):358-373

[9] ECRI Institute (2012) Top 10 health technology hazards for 2013. Health Devices 41(11):342-365

[10] ECRI Institute (2013) Top 10 heath technology hazards for 2014. Health Devices 42(11):354-380

[11] ECRI Institute (2014) Top 10 heath technology hazards for 2015. Health Devices

[12] ECRI Institute (2015) Top 10 heath technology hazards for 2016. Health Devices

[13] ECRI Institute (2016) Top 10 heath technology hazards for 2017. Health Devices

[14] Clifford GD, Silva I, Moody B, Li Q, Kella D, Shahin A et al (2015) The PhysioNet/Computing in Cardiology Challenge 2015: reducing false arrhythmia alarms in the ICU. Comput Cardiol 2015 273-276

[15] Eerikäinen, LM, Vanschoren J, Rooijakkers MJ, Vullings R, Aarts RM (2016) Reduction of false arrhythmia alarms using signal selection and machine learning. Physiol Meas 37(8):204

[16] Rijsbergen CJ (1979) Information Retrieval, 2nd edn. London: Butterworths

[17] Bitan,Y, O' Connor M F (2012) Correlating data from different sensors to increase the positive predictive value of alarms: an empiric assessment. F1000research 1:45

[18] Imhoff M, Kuhls S (2006) Alarm algorithms in critical care monitoring. Anesth Analg 102(5):1525-1537

[19] Vesin A, Azoulay E, Ruckly S, Vignoud L, Rusinovà K, Benoit D, et al (2013) Reporting and handling missing values in clinical studies in intensive care units. Intensive Care Med 39(8):1396-1404

[20] Altman DG, Bland JM (1994) Statistics notes: diagnostic tests 2: predictive values. Br Med J 30(6947):102

[21] Lalkhen AG, McCluskey A (2008) Clinical tests: sensitivity and specificity. Contin Educ Anaesth Crit Care Pain 8(6):221-223

[22] Božikov J, Zaletel-Kragelj L (2010) Test validity measures and receiver operating characteristic (ROC) analysis. Methods Tools Public Health 749-770

[23] Bishop CM (2007) Pattern recognition and machine learning. New York: Springer

[24] Liaw A, Wiener M (2002) Classification and regression by random forest. R news 2(3):18-22

[25] Breiman L (1984) Classification and regression trees. Belmont, Calif: Wadsworth International Group

Appendices

Appendix A – Full results for the first test

Note: Unlike the RF model for which the output can be thresholded with different values, the FER/PER classification results are a single value (a binary output obtained by the model rules, e.g., ARTBPM < 50 mm Hg, CVP < 5 mm Hg, and CVP > -10 mm Hg indicate deterministically Hypovolemia). Thus, it is impossible to calculate the value of the area under the curve (AUC) for PER, and therefore these cells below are left empty. Green and orange cells indicate better and worse results, respectively, for each of the missing parameters in each clinical alarm scenario.

<u>One missing parameter</u>:

Missing parameter	Model	Youden's index	PPV	Specificity	Sensitivity	FAR	AUC
HR	RF	0.94	0.99	0.99	0.95	0.01	0.99
	PER	0.72	0.84	0.84	0.89	0.16	
ARTBPS	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.22	0.56	0.22	1	0.78	
ARTBPM	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.89	0.94	0.94	0.95	0.06	
CVP	RF	0.99	1	1	0.99	0	1
	PER	0.99	1	1	0.99	0	
PAPD	RF	0.99	1	1	0.99	0	1
	PER	0.93	0.94	0.94	0.99	0.06	
RR total	RF	0.99	1	1	0.99	0	1
	PER	0.93	0.94	0.94	0.99	0.06	
RR mandatory	RF	0.99	1	1	0.99	0	1
	PER	0.93	0.94	0.94	0.99	0.06	
Average	**RF**	**0.97**	**0.99**	**0.97**	**0.98**	**0.0042**	**0.99**
	PER	**0.80**	**0.88**	**0.83**	**0.97**	**0.1687**	

Two missing parameters:

Missing parameters	Model	Youden's index	PPV	Specificity	Sensitivity	FAR	AUC
HR + ARTBPS	RF	0.91	0.97	0.97	0.93	0.03	0.99
	PER	0.06	0.51	0.1	0.95	0.9	
HR + ARTBPM	RF	0.93	0.98	0.98	0.95	0.02	0.99
	PER	0.69	0.84	0.84	0.86	0.16	
HR + CVP	RF	0.94	0.98	0.98	0.95	0.02	0.99
	PER	0.77	0.89	0.89	0.87	0.11	
HR + PAPD	RF	0.94	0.98	0.98	0.95	0.02	0.99
	PER	0.72	0.84	0.84	0.89	0.16	
HR + RR total	RF	0.94	0.98	0.99	0.95	0.01	0.99
	PER	0.72	0.84	0.84	0.89	0.16	
HR + RR mandatory	RF	0.94	0.98	0.99	0.95	0.01	0.99
	PER	0.72	0.84	0.84	0.89	0.16	
ARTBPS + ARTBPM	RF	0.91	0.97	0.97	0.94	0.03	0.99
	PER	0.2	0.55	0.22	0.97	0.78	
ARTBPS + CVP	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.26	0.57	0.27	1	0.73	
ARTBPS + PAPD	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.01	0.5	0.01	1	0.99	
ARTBPS + RR total	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.22	0.56	0.22	1	0.78	
ARTBPS + RR mandatory	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.22	0.56	0.22	1	0.78	
ARTBPM + CVP	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.9	1	1	0.9	0	
ARTBPM + PAPD	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.9	0.94	0.94	0.96	0.06	
ARTBPM + RR total	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.89	0.94	0.94	0.95	0.06	
ARTBPM + RR mandatory	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.89	0.94	0.94	0.95	0.06	
CVP + PAPD	RF	0.99	1	1	0.99	0	1
	PER	0.99	1	1	0.99	0	
CVP + RR total	RF	0.99	1	1	0.99	0	1
	PER	0.99	1	1	0.99	0	
CVP + RR mandatory	RF	0.99	1	1	0.99	0	1
	PER	0.99	1	1	0.99	0	
PAPD + RR total	RF	0.99	1	1	0.99	0	1
	PER	0.93	0.94	0.94	0.99	0.06	
PAPD + RR mandatory	RF	0.99	1	1	0.99	0	1
	PER	0.93	0.94	0.94	0.99	0.06	
RR total + RR mandatory	RF	0.99	1	1	0.99	0	1
	PER	0.93	0.94	0.94	0.99	0.06	
Average	**RF**	**0.96**	**0.98**	**0.98**	**0.97**	**0.0104**	**0.99**
	PER	**0.66**	**0.81**	**0.71**	**0.95**	**0.289**	

Three missing parameters:

Missing parameters	Model	Youden's index	PPV	Specificity	Sensitivity	FAR	AUC
HR + ARTBPS + ARTBPM	RF	0.83	0.93	0.94	0.89	0.06	0.97
	PER	0.04	0.51	0.1	0.93	0.9	
HR + ARTBPS + CVP	RF	0.9	0.97	0.97	0.93	0.03	0.99
	PER	0.06	0.51	0.13	0.94	0.87	
HR + ARTBPS + PAPD	RF	0.9	0.97	0.97	0.93	0.03	0.99
	PER	0	0.5	0.01	0.99	0.99	
HR + ARTBPS + RR total	RF	0.9	0.97	0.97	0.93	0.03	0.99
	PER	0.13	0.54	0.24	0.89	0.76	
HR + ARTBPS + RR mandatory	RF	0.9	0.97	0.97	0.93	0.03	0.99
	PER	0.12	0.53	0.23	0.89	0.77	
HR + ARTBPM + CVP	RF	0.92	0.97	0.98	0.95	0.02	0.99
	PER	0.67	0.88	0.89	0.78	0.11	
HR + ARTBPM + PAPD	RF	0.92	0.98	0.98	0.95	0.02	0.99
	PER	0.69	0.84	0.84	0.86	0.16	
HR + ARTBPM + RR total	RF	0.93	0.98	0.98	0.95	0.02	0.99
	PER	0.69	0.84	0.84	0.86	0.16	
HR + ARTBPM + RR mandatory	RF	0.92	0.98	0.98	0.95	0.02	0.99
	PER	0.69	0.84	0.84	0.86	0.16	
HR + CVP + PAPD	RF	0.93	0.98	0.98	0.95	0.02	0.99
	PER	0.77	0.89	0.89	0.87	0.11	
HR + CVP + RR total	RF	0.93	0.98	0.98	0.95	0.02	0.99
	PER	0.77	0.89	0.89	0.87	0.11	
HR + CVP + RR mandatory	RF	0.94	0.98	0.98	0.95	0.02	0.99
	PER	0.77	0.89	0.89	0.87	0.11	
HR + PAPD + RR total	RF	0.94	0.98	0.98	0.95	0.02	0.99
	PER	0.72	0.84	0.84	0.89	0.16	
HR + PAPD + RR mandatory	RF	0.93	0.98	0.98	0.95	0.02	0.99
	PER	0.72	0.84	0.84	0.89	0.16	
HR + RR total + RR mandatory	RF	0.94	0.98	0.98	0.95	0.02	0.99
	PER	0.72	0.84	0.84	0.89	0.16	
ARTBPS + ARTBPM + CVP	RF	0.89	0.96	0.97	0.92	0.03	0.99
	PER	0.15	0.54	0.27	0.88	0.73	
ARTBPS + ARTBPM + PAPD	RF	0.9	0.96	0.96	0.93	0.04	0.99
	PER	0.01	0.5	0.01	1	0.99	
ARTBPS + ARTBPM + RR total	RF	0.9	0.96	0.97	0.93	0.03	0.99
	PER	0.2	0.55	0.22	0.97	0.78	
ARTBPS + ARTBPM + RR mandatory	RF	0.9	0.96	0.97	0.93	0.03	0.99
	PER	0.2	0.55	0.22	0.97	0.78	
ARTBPS + CVP + PAPD	RF	0.96	0.99	0.99	0.98	0.01	1
	PER	0.76	0.9	0.9	0.86	0.1	
ARTBPS + CVP + RR total	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.26	0.57	0.27	1	0.73	

ARTBPS + CVP + RR mandatory	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.24	0.57	0.29	0.96	0.71	
ARTBPS + PAPD + RR total	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.01	0.5	0.01	1	0.99	
ARTBPS + PAPD + RR mandatory	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.01	0.5	0.01	1	0.99	
ARTBPS + RR total + RR mandatory	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.22	0.56	0.22	1	0.78	
ARTBPM + CVP + PAPD	RF	0.97	0.99	0.99	0.98	0.01	1
	PER	0.9	1	1	0.9	0	
ARTBPM + CVP + RR total	RF	0.98	0.99	0.99	0.98	0.01	1
	PER	0.9	1	1	0.9	0	
ARTBPM + CVP + RR mandatory	RF	0.98	0.99	0.99	0.98	0.01	1
	PER	0.9	1	1	0.9	0	
ARTBPM + PAPD + RR total	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.9	0.94	0.94	0.96	0.06	
ARTBPM + PAPD + RR mandatory	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.9	0.94	0.94	0.96	0.06	
ARTBPM + RR total + RR mandatory	RF	0.98	0.99	0.99	0.99	0.01	1
	PER	0.89	0.94	0.94	0.95	0.06	
CVP + PAPD + RR total	RF	0.99	1	1	0.99	0	1
	PER	0.99	1	1	0.99	0	
CVP + PAPD + RR mandatory	RF	0.99	1	1	0.99	0	1
	PER	0.99	1	1	0.99	0	
CVP + RR total + RR mandatory	RF	0.99	1	1	0.99	0	1
	PER	0.99	1	1	0.99	0	
PAPD + RR total + RR mandatory	RF	0.99	1	1	0.99	0	1
	PER	0.93	0.94	0.94	0.99	0.06	
Average	**RF**	**0.94**	**0.98**	**0.98**	**0.95**	**0.018**	**0.99**
	PER	**0.54**	**0.76**	**0.61**	**0.92**	**0.386**	