



BEN-GURION UNIVERSITY OF THE NEGEV  
FACULTY OF ENGINEERING SCIENCES  
DEPARTMENT OF INDUSTRIAL ENGINEERING AND MANAGEMENT

**Machine learning applied multi-sensor information to  
reduce the impact of missing data and false alarms of  
intensive care unit monitors**

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE M.Sc DEGREE

By: **Gal Hever**

Supervised by: **Dr. Yuval Bitan**

October 2017



BEN-GURION UNIVERSITY OF THE NEGEV  
FACULTY OF ENGINEERING SCIENCES  
DEPARTMENT OF INDUSTRIAL ENGINEERING AND MANAGEMENT

**Machine learning applied multi-sensor information to  
reduce the impact of missing data and false alarms of  
intensive care unit monitors**

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE M.Sc DEGREE

By: Gal Hever

Supervised by: Dr. Yuval Bitan


October 2017

Author: Gal Hever

.....  .....

Date: 26-10-2017

Supervisor: Dr. Yuval Bitan

.....  .....

Date: 26-10-2017

Chairman of Graduate Studies Committee: Prof. Israel Parmet

.....  ..... Date: 26-10-2017

# Abstract

**Background:** While monitoring a patient's clinical state in intensive care units (ICUs), clinical alarms have become an indispensable part of the medical environment. However, multiple studies reveal that the false alarm rates (FAR) of ICU monitors range from 72% to 99%, which negatively affect both patients and medical staff. This research apply machine-learning (ML) methods to ICU multi-sensor information to imitate a medical specialist in diagnosing a patient's condition. The system was trained to reduce the FAR without compromising the patient's safety in critical care even when sensor information is missing.

**Methods:** A large multi-parameter patient database was established at the Tel-Aviv Medical Center (TAMC) and used to identify based on expert rules an alarm set for seven clinical scenarios: Bradycardia, Bradycardia hypotension, Hypovolemia, Tachycardia, Tachycardia hypotension, Obstructive shock, and Left ventricular (LV) shock. Each data sample portrays simultaneous measuring at a specific minute of twelve physiological parameters (such as, heart rate and different blood pressures) based on information from several sensors and was tagged as “alarm” or “not alarm” by a clinical expert based on the seven rules. Since it is often the case that not all the sensors are available, a random forest (RF) ML model was trained and evaluated on datasets with increasing numbers of missing parameters compared with the full expert-based rules (FER), which is the ground truth, and partial expert-based rules (PER), in which rules that contained the missing parameters were removed. FER represents a case in which all sensors are available for a decision about a patient by the medical specialist, whereas PER is the case where some of these sensors are unavailable (broken or do not exist at a specific ICU/time). RF, FER, and PER models were evaluated using the conventional precision, specificity, and sensitivity performance measures and also using the Youden’s statistic (index), which evaluates a decision taken based on the two latter measures simultaneously.

**Results:** In the first test, all seven clinical alarm scenarios were examined collectively, while in the second test, each scenario was examined separately. In the first test, the absence of the heart rate (HR), arterial blood pressure systolic (ARTBPS), arterial blood pressure mean (ARTBPM), and pulmonary artery pressure diastolic (PAPD)

parameters has led to the poorest performances of RF and PER. In this test, in the absence of the parameters, the Youden's index and precision measure for RF varied between 0.94 and 0.99 and 0.98 and 0.99, respectively, in comparison to PER for which these measures varied between 0.54 and 1 and 0.76 and 1, respectively. In the second test, the Youden's index and precision measure varied between 0.43 and 1 and 0.65 and 1 for RF and 0.25 and 1 and 0.58 and 1 for PER.

**Conclusions:** If all sensors in an ICU are available for making a decision about a patient, a trained-from-data RF ML model is equally accurate as a medical specialist having years of experience in the ICU. However, while the RF model attains its high accuracy and low FAR when sensor data becomes missing, the specialist performance worsens with the number of missing sensors. For example, decisions made based on the specialist rules incur FAR of 6%, 11%, or 23% when one, two, or three sensors are missing, respectively, while those based on the RF lead to a steady FAR of only 2-3%. RF success is attributed to its ability to integrate and fuse information from the available sensors compensating for those missing, demonstrating that the ICU information processing human mechanism is probably less efficient than that of the proposed ML-based approach, soliciting an ML-based ICU decision support system.

**Keywords:** Intensive Care Unit, False Alarms, Machine Learning, Random Forests

# Acknowledgement

First and foremost, I would sincerely thank my thesis advisor, Dr. Yuval Bitan for his dedicated guidance throughout the research. I am truly grateful for your help and support in my endeavors. Thank you for sharing with me your ideas, experience, time, and understanding.

I would also like to express my sincere thanks to Prof. Boaz Lerner for his much-appreciated support and guidance throughout my thesis.

This research was done in cooperation with the Intensive Care Unit Department of Tel-Aviv Medical Center. I would like to express my appreciation and thanks to Prof. Idit Matot, The Chair at Division of Anesthesia at Pain Critical Care, Tel-Aviv Medical Center, who provided the clinical data. Sincere thanks to the other unit members involved in this research especially Hanna Artsi, for her advice, support and time greatly enhancing the shaping of this study.

I am also grateful to Prof. Michael F. O'Connor, Section Head of Anesthesia & Critical Care Department at The University of Chicago, for sharing his time, insights and knowledge of clinical decisions which was critical to the success of this research.

I also want to give special thanks to Eran Turgeman and all the rest of the IT team, for providing the work environment and technical support needed for conducting this work.

I will take this opportunity to thank Tamar Domany from Intensix Company for her great help with the medical database.

Last but not least, I express my deep appreciation to my supportive family and especially to my boyfriend for his incredible understanding, tolerance and love during my research.

# Table of contents

|  |    |
|--|----|
| ABSTRACT .....   | 3  |
| ACKNOWLEDGEMENT .....  | 5  |
| TABLE OF CONTENTS.....   | 6  |
| LIST OF TABLES.....  | 8  |
| LIST OF FIGURES.....   | 9  |
| LIST OF ACRONYMS AND ABBREVIATIONS .....                                       | 10 |
| 1. INTRODUCTION .....  | 11 |
| 2. THEORETICAL BACKGROUND REVIEW .....   | 12 |
| 2.1. SIGNAL DETECTION THEORY .....   | 12 |
| 2.1.1. <i>The Fundamental Decision Problem</i> .....                           | 12 |
| 2.1.2. <i>The Receiver Operating Characteristic Curve</i> .....                | 15 |
| 2.1.3. <i>Response Criterion</i> .....   | 16 |
| 2.1.4. <i>The discriminability index</i> .....                                 | 16 |
| 2.1.5. <i>SDT tools in Diagnostic Medicine</i> .....                           | 17 |
| 2.2. MEDICAL MONITORING .....  | 19 |
| 2.2.1. <i>Patient Monitoring</i> .....   | 19 |
| 2.2.2. <i>Alarm designation</i> .....  | 21 |
| 2.2.3. <i>False Alarms</i> .....   | 22 |
| 2.3. DECISION TREES AND FORESTS .....  | 28 |
| 2.3.1. <i>Machine Learning overview</i> .....                                  | 28 |
| 2.3.2. <i>Decision Trees</i> .....   | 28 |
| 2.3.3. <i>Attribute Selection Measures</i> .....                               | 31 |
| 2.3.4. <i>Random Forests</i> .....   | 33 |
| 2.4. TOOLS FOR DIAGNOSTIC TESTS PERFORMANCE EVALUATION .....                   | 35 |
| 2.4.1. <i>ROC curve as an Assessment Measurement in Machine Learning</i> ..... | 36 |
| 2.4.2. <i>AUC</i> .....  | 36 |
| 2.4.3. <i>Confusion matrix</i> .....   | 37 |
| 2.4.4. <i>Youden's index</i> .....   | 38 |
| 2.5. LITERATURE REVIEW SUMMARY .....   | 40 |
| 3. RESEARCH METHODOLOGY .....  | 41 |
| 3.1. MODEL OVERVIEW .....  | 41 |
| 3.2. DATA RESOURCE .....   | 42 |
| 3.3. DATA UNDERSTANDING .....  | 44 |
| 3.3.1. <i>Independent Variables - Medical Monitoring Parameters</i> .....      | 44 |
| 3.3.2. <i>Dependent variable</i> .....   | 46 |
| 3.3.3. <i>Clinical Alarm Scenarios</i> .....                                   | 46 |
| 3.3.4. <i>Description statistics</i> .....                                     | 48 |
| 3.4. DATA PREPARATION .....  | 53 |
| 3.5. EXPERT-BASED RULES IMPLEMENTATION .....                                   | 53 |
| 3.6. EXPERT-BASED RULES EVALUATION.....  | 54 |

|   |    |
|---|----|
| 3.7. METHODOLOGY IMPLEMENTATION .....                                   | 56 |
| 4. RESULTS.....   | 59 |
| 5. DISCUSSION .....   | 66 |
| 6. CONCLUSIONS .....  | 67 |
| 7. MERIT OF THE RESEARCH AND PROPOSED CONTRIBUTION TO SCIENCE .....     | 68 |
| 8. LIMITATIONS AND FUTURE RESEARCH .....                                | 69 |
| 9. APPENDIXES.....  | 70 |
| 9.1. APPENDIX A – CLASSIFICATION TERMS.....                             | 70 |
| 9.2. APPENDIX B – DESCRIPTIVE STATISTICS .....                          | 70 |
| 9.3. APPENDIX C – ALARM FREQUENCY .....                                 | 71 |
| 9.4. APPENDIX D – FER PERFORMANCES IN CASES OF MISSING PARAMETERS ..... | 72 |
| 9.5. APPENDIX E – FULL RESULTS .....                                    | 74 |
| REFERENCES.....   | 79 |
| <i>Academic References</i> .....  | 79 |
| תקציר.....  | 84 |

## List of tables

|  |    |
|--|----|
| Table 1 - Response matrix for SDT that maps the state-of the-world to diagnostic system response. Adapted from Green and Swets (1966). .....   | 14 |
| Table 2 - confusion matrix for a binary classifier. ....   | 37 |
| Table 3 - Default alarm settings. ....   | 43 |
| Table 4 - Normal ranges of parameters for an adult. ....   | 46 |
| Table 5 - Explained variable.....  | 46 |
| Table 6 - Monitor alarms per shift.....  | 51 |
| Table 7 - Monitor alarms per hour. ....  | 51 |
| Table 8 - Parameter frequency and missingness in the TAMC database. ....   | 52 |
| Table 9 - The full expert-based rules that make the research's ground truth. ....  | 54 |
| Table 10 - Percentages of monitor alarms. ....   | 55 |
| Table 11 - Percentages of FER alarms. ....   | 55 |
| Table 12 - Confusion matrix of monitor performances. ....  | 55 |
| Table 13 - Average models performance with and without missing parameters. Note that PER for no missing data is the FER, which is the reference/benchmark result of this research.....   | 59 |
| Table 14 - Averages model performance measuresfor different clinical alarm scenario and missing parameter setting. Green/red cells indicate higher/lower scores in comparing RF and PER for each of the missing parameters in each clinical alarm scenario. An empty cell indicates the inability to exercise the PER model due to missing parameter(s) this model is based on. .... | 61 |
| Table 15 - Parameter importance ranking according to the Gini index.....   | 64 |



## List of figures

|  |    |
|--|----|
| Figure 1 - Representation of the Signal Detection Theory decision space. The x-axis represents the magnitude of the criterion (Higham & Arnold ,2007).....   | 15 |
| Figure 2 - Physiologic monitoring devices in the ICU (Drew et al., 2014).....  | 20 |
| Figure 3 - A decision tree for concept Play Tennis, an example is classified by sorting it through the tree to the appropriate leaf node, then returning the classification associated with leaf (in this case Yes or No). This tree classifies Saturday morning (Mitchell, 1997)..... | 30 |
| Figure 4 - Finding best cut-off from the ROC curve by using Youden's index J (Božikov & Zaletel-Kragelj, 2010). ....   | 38 |
| Figure 5 - Histogram Tab in the data visualization tool. ....  | 48 |
| Figure 6 - Classification feature in the visualization tool.....   | 49 |
| Figure 7 - Total daily monitor alarm (clinical and technical) distribution among day and hour. ....  | 50 |
| Figure 8 - Average per patient of daily monitor alarm (clinical and technical) distribution among day and hour. ....   | 50 |
| Figure 9 – FER' performance in cases of missing parameters.....  | 56 |
| Figure 10 - Models' Youden's indexes for missing parameters.....   | 60 |

## List of acronyms and abbreviations

| <b>Term</b>  | <b>Description</b>                     |
|--------------|--|
| ARTBPM       | Arterial Blood Pressure Mean           |
| ARTBPS       | Arterial Blood Pressure Systolic       |
| CVP          | Central Venous Pressure                |
| ECRI         | Emergency Care Research Institute      |
| FA           | False Alarms                           |
| FAR          | False Alarms Rate                      |
| FER          | Full Expert-Based Rules                |
| Fio2         | Fraction of Inspired Oxygen            |
| HR           | Heart Rate                             |
| ICU          | Intensive Care Unit                    |
| NPV          | Negative Predictive Value              |
| NTP          | Nurse to Patient Ratio                 |
| PAPD         | Pulmonary Artery Pressure Diastolic    |
| PER          | Partial Expert-Based Rules             |
| PPV          | Positive Predictive Value              |
| RF           | Random Forest                          |
| ROC          | Receiver Operating Characteristic      |
| RR Mandatory | Mechanical Respiratory Rate            |
| RR Total     | Total Respiratory Rate                 |
| SDT          | Signal Detection Theory                |
| Spo2         | Saturation Peripheral Capillary Oxygen |
| TAMC         | Tel-Aviv Medical Center                |

# 1. Introduction

This research aims at developing automated data-driven decision support tools using machine-learning (ML) methodologies towards minimizing the false alarm rate (FAR) of intensive care unit (ICU) monitors. The present work is organized as follows:

- Chapter 2 extends the discussion of Signal Detection Theory (SDT) and offers different approaches for data analysis and interpretation. The review presents previous studies that discuss ICU monitoring, including its existing situation and work configuration, critical problems and solutions. The literature review continues with a general introduction of ML focused on random forest (RF) model. Finally, relevant tools for diagnostic test performances are demonstrated.
- Chapter 3 presents a description about a large-scale, comprehensive medical database that includes physiological and clinical data collected from mature patients admitted to the ICU Department in Tel-Aviv Medical Center (TAMC). Then, the methodology framework of this work is presented which includes an integration between an expert-based method and a ML ensemble model. Essential descriptive statistics of the data in some visualization dimensions as part of the medical analysis and a prerequisite for understanding further statistical evaluations are also demonstrated in the end of this chapter.
- Chapter 4 demonstrates the classification results and advances of the adjusted RFs model based on the full expert-based rules (FER) in predicting alarm events and reducing alarm frequency.
- Chapter 5 describes the findings in the form of a set of tables and graphs and discuss the results of the statistical analysis.
- Chapter 6 draws conclusions.
- Chapter 7 summarizes the research contributions in the present thesis.
- Finally, Chapter 8 discusses limitations, and several research ideas for enhancing patient monitoring and safety, along with quality of care rooted in the FER proposed in this research.

## **2. Theoretical Background Review**

### **2.1. Signal Detection Theory**

Signal Detection Theory (SDT) that was developed by Tanner & Swets in 1954, is a statistical model used originally to detect aircraft signals on radar displays for military needs. SDT has proven to be a useful and robust model that applied problems related to detection of stimuli in noise (random patterns). Later the model was extended to a variety of research domains and proposed as a method for prediction quality evaluation (Spackman, 1989). Statistical classification is needed in many fields from computer science, economics, meteorology, biology, biochemistry and medical studies. For example, auditory signals in communication systems (Peterson, Birdsall & Fox, 1954) while the detected psychophysical stimulus is referred to as the signal and all other environmental events are defined as noise.

This model employs wide-ranging techniques to quantify the ability of a detection system (whether it be a human, a test or a device) to distinguish between signal and noise. This chapter will provide an introduction to these techniques and will explain the theoretical principles of SDT model in the context of medical decision-making in intensive care units (ICUs).

#### **2.1.1. The Fundamental Decision Problem**

The fundamental decision problem includes an environment that involves the combination of stimulus and response. Where one or more stimuli occurred, the diagnostic system had to determine which stimulus occurred and respond in accordance (Swets, Tanner & Birdsall, 1961). The combination of stimulus and response is called an event and is comprised of three components (Green & Swets, 1966):

- The state-of-the-world (essentially the ground truth).
- The information the system receives.

- The decision related to the classification of the stimulus as a signal or noise event.

SDT maintains that response to a given stimulus will be different for each diagnostic system and depends upon two factors: the sensitivity of the system and the response criterion. The system responds affirmatively to signal presence, if the stimulus magnitude is larger than an internal value (decision criterion) and will respond negatively if the stimulus magnitude is less than the criterion (Peterson, Birdsall & Fox, 1954).

The information the diagnostic system receives for making the decision contains noise, and the system's ability to filter out true signals from the constant background noise is called detection. SDT is a means to quantify the ability to discern between a signal and the noise (Egan, 1975).

There are four possible outcomes resulting from the combination of the two states-of-the-world and the two response alternatives: Hit, False alarm (FA), Miss and Correct Rejection. The diagnostic system might make errors when the signal is very faint, or the noise level is very high. These two types of errors are called false positives and false negatives.

False positives occur if a system response is positive but a signal has not been presented. This event is termed a FA. For example, in medicine FA is a falsely message that disease is present, when it is actually absent.

$$\text{False alarm rate} = \frac{\text{Number of false alarms}}{\text{Number of noise trials}} \quad (1)$$

False negatives (FN) occur when the response of a system is negative but actually a signal was presented, this event is termed a Miss. For example, in the medical world FN is a falsely reassuring message to patients and physicians that disease is absent, when it is actually present.

When the system response is positive and the signal is presented, the event is termed a Hit.

$$\text{Hits rate} = \frac{\text{Number of hits}}{\text{Number of signal trials}} \quad (2)$$

In the last possible event of responding negatively when no signal has been presented, the event is termed a Correct Rejection (Green & Swets, 1966).

The response matrix below (Table 1) summarized all the four possible events that can be computed. Hits and Correct Rejections are both correct responses. Misses and FA are both incorrect responses. Hits and Misses are related, and Correct Rejections and FA are related, therefore, researchers report only Hits and FA. (The proportion of Misses is  $(1 - \text{Hits})$ ; the proportion of Correct Rejections is  $(1 - \text{FA})$ ).

| <div style="text-align: center;"> <b>Response<br/>Alternatives</b><br/><br/> <b>State of the World<br/>Alternatives (Stimulus)</b> </div> |                            |                                  |
|---|----------------------------|----------------------------------|
|   | S (“Yes”)                  | N (“No”)                         |
| Signal Present (s)  | P(S s)<br>Hit (H)          | P(N s)<br>Miss (M)               |
| Signal Absent/Noise (n)   | P(S n)<br>False alarm (FA) | P(N n)<br>Correct Rejection (CR) |

*Table 1 - Response matrix for SDT that maps the state-of-the-world to diagnostic system response. Adapted from Green and Swets (1966).*

S = Affirmative response regarding signal presence.

N = Negative response regarding signal presence.

s = The signal is present.

n = The signal is absent ('noise' or non-signal).

### 2.1.2. The Receiver Operating Characteristic Curve

Receiver Operating Characteristic (ROC) curve is a detection model that describes the relationship between the Hit (true positive) rate against the FA (false positive) rate as the diagnostic system changes decision criteria (various threshold settings) (Green & Swets, 1966).

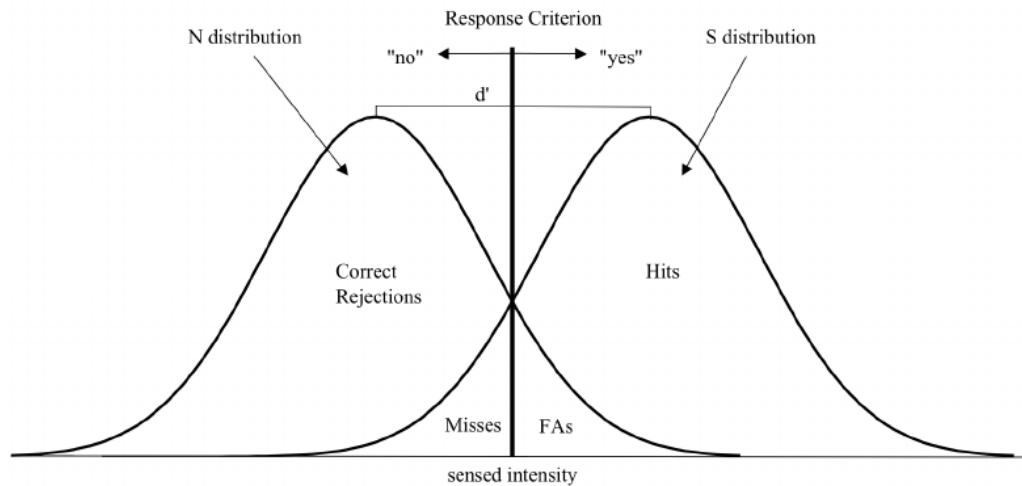


Figure 1 - Representation of the Signal Detection Theory decision space. The x-axis represents the magnitude of the criterion (Higham & Arnold ,2007).

This curve (Figure 1) consists of two normal distributions, one representing a signal and the other one representing the noise (signal absent). The curve will bow upward in term of any reasonable choice of criterion (Hit rate is always larger than FA) and illustrates various criteria, given a signal and noise distribution (Higham & Arnold ,2007).

Two factors effect on a system's perception of a given stimulus according to SDT:

- a) **Response Criterion** – the location of the measure of bias.
- b) **Discriminability Index** - the sensitivity of the system.

(Egan, 1975; Swets, Dawes & Monahan, 2000)

### 2.1.3. Response Criterion

Response criterion is the location of the decision criterion that is called  $\beta$  (measure of bias). Bias is the extent to which one response is more probable than another. This tendency is affected by factors independent of the intensity of the stimuli. The system's decision depends on the magnitude of the measured value. The system decides that a signal is present if the magnitude of x-axis is greater than the criteria and vice versa (Gescheider, 1997).

According to Figure 1, criterion line divides the graph into four sections (Hits, Misses, FA and Correct Rejections). The criterion can be described as low, unbiased, or high. A low criterion would indicate a lower magnitude value of the stimulus for the system to respond in the affirmative to the signal presence, which will cause a positive response to almost every event. In this situation, any event will never be missed, which results in more Hits but at the cost of more FA. A high criterion would indicate a higher magnitude of the stimulus that is required respond in the affirmative regarding signal presence. It will cause a negative response to almost every event, thus more events will be missed. These events result in fewer FA, but at the cost of fewer correct detections (Higham & Arnold, 2007).

### 2.1.4. The discriminability index

The discriminability index is a measure of sensitivity of the sensory process that is also called  $d'$ . This index estimates how well a diagnostic system can discriminate between signal present and signal absent trials. This measure is represented by the difference between the means output of the two distributions. The higher the signal detection performance the better the system's ability to truly discriminate between signal and noise. This stems from the fact that larger  $d'$  value will have higher sensitivity and therefore will have a higher rate of Correct Rejections and a lower rate of FA (Macmillan & Creelman, 2005).

$$d' = Z(\text{hit rate}) - Z(\text{false alarm rate}) \quad (3)$$



### **2.1.5. SDT tools in Diagnostic Medicine**

Numerous studies in medical diagnosis have relied on SDT concepts. For example, ROC curves are widely used to evaluate diagnostic accuracy of a laboratory test (Zweig & Campbell, 1993; Pepe, 2003).

In the field of medicine analysis, the stakes involve life-threatening situations and the diagnosis of the status of a subject is crucial to its accurate classification. That, combined with time pressure, can generate considerable stress for the physician or caregiver making the forecasts and detections. Rigorous evaluation of physician judgments and practices using the methods proposed in this chapter have improved medical outcomes substantially (McFall & Treat 1999; Lusted, 1971).

The fundamental medical decision problem includes an environment that allows doctors to use their own judgement. For example, examining a CT scan combines level of uncertainty, either there is a tumor (signal present) or not (signal absent). There are four possible outcomes: Hit (tumor present and doctor says "yes"), Miss (tumor present and doctor says "no"), false alarm (tumor absent and doctor says "yes"), and Correct Rejection (tumor absent and doctor says "no"). Some doctors may choose to be more conservative and say "no tumor" more often. They will miss more tumors, but they will be doing their part to reduce unnecessary surgeries. Two doctors, with equally training, looking at the same CT scan, will judge the same situation differently because each may have a different response criterion.

This chapter reviews a tool that helps to make decisions under uncertainty conditions and quantify the ability to discern between signal and noise. This work apply an algorithm that correlates information across sensors that are used to detect and suppress artifact in a manner similar to how human operators analyze data. The FER that was developed in this thesis and will be presented in following sections uses predefined rules to alert the medical staff to abnormal situations. The concept of the FER is similar to the signal to noise ratio reviewed in this chapter.

Moreover, STD tools have been borrowed to classify algorithms from machine-learning (ML) domain for algorithm evaluation. This research have used these tools for assessing the ML algorithms used for dealing with missing monitor's sensors. The combination between these two worlds will be presented in the following chapters.

## **2.2. Medical Monitoring**

### **2.2.1. Patient Monitoring**

Monitoring patient status in hospitals is indispensable especially in ICUs and requires adequate medical devices. These devices support the staff in assessing a patient's health status. Monitoring by devices can also compensate for the lack of manpower or suboptimal staffing levels in ICUs (DeVita et al., 2010). These monitor devices have an automatic alarm system to alert the healthcare professionals or the caregivers to abnormalities in patient's vital signs so that timely clinical decisions can be made to prevent complications (Imhoff & Kuhls, 2006).

Today most of the alarms are univariate in practice and typically triggered independently when any of the individual parameters has crossed a predefined "low" or "high" threshold range for a few seconds (physiological state of the patient needs attention). Additionally, the alarm will be triggered when a technical issue occurs (Drew et al., 2014). When a threshold is reached, the triggered alarms could be presented audibly, visually, and through vibration in a mobile device depending on its prioritization, adapting risk to an appropriate level of alertness (Cvach, 2012).

An important aspect of patient safety in monitoring is the sensitivity and specificity of the alarms. Sensitivity and specificity are terms used to evaluate classification accuracy. The algorithm employed by the monitors has a standard definition of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). These terms are used for evaluating the clinical test of the monitor and can be found in Appendix A (Lalkhen & McCluskey, 2008). These terms will be reviewed in the next paragraphs.



*Figure 2 - Physiologic monitoring devices in the ICU (Drew et al., 2014)*

Sensitivity refers to the ability of the monitor to correctly identify those patients with the problem or in case of this study, identify samples that should cause alarm (Lalkhen & McCluskey, 2008). For example, a monitor with 80% sensitivity detects 80% of alarms (true positives) but 20% go undetected (false negatives). A high sensitivity is clearly important where the monitor is used to identify a patient in a life-threatening condition. The sensitivity of a perfect monitor would be 100%, never missing a clinically important event.

Specificity refers to the ability to correctly identify those patients without the problem (Lalkhen & McCluskey, 2008) or the samples that should be classified as no-alarm in case of this study. For example, a monitor with 80% specificity correctly reports 80% of no-alarm (true negatives) but 20% of no-alarm are incorrectly identified (false positives). Specificity would be 100% when the

alarm doesn't go off while there is no clinically important event (Lalkhen & McCluskey, 2008).

In theory a perfect alarm system would never miss a clinically important event—the sensitivity would be 100%. Also, the alarm would not go off when there is no clinically important event—the specificity would be 100%. However, in reality diagnostic tests are never perfect. The higher the sensitivity, the lower the specificity, and vice versa (Sendelbach & Funk, 2013).

PPV is another measure for classification results that is also known as precision. The PPV of a medical monitor describes the percentage of patients with a positive test who actually have the problem. This term is useful since it answers the question: 'When the monitor predicts an alarm, how often is it correct?'

NPV of a medical monitor describes the percentage of patients with a negative test who do not have a problem. It can answer the question: 'When the monitor didn't predict an alarm, how often is it correct?' (Lalkhen & McCluskey, 2008).

### **2.2.2. Alarm designation**

Alarms are expected to increase patient safety by early detection of any abnormality in patient condition and essential device function (Drew et al., 2014). The monitor has several goals which are hierarchically ranked according to importance. The main goal is detection of life-threatening situations such as asystole or extremely low blood pressures. The second goal is detection of device malfunction that can lead to a life-threatening event such as disconnection from the patient or power source. Third goal is detection of imminent danger before it becomes life-threatening event. Another goal is detection of imminent device malfunction that warns the caregiver before the device crashes. The last goal is diagnostic pathophysiologic condition (for instance, hypovolemia) rather than drawing attention to the out of range

parameters. The overarching purpose is to detect changes early and suggest appropriate remedial treatment (Imhoff & Kuhls, 2006).

### **2.2.3. False Alarms**

Alarms have an important role in man-machine interface. However, several significant shortcomings concerning those existing bedside physiologic monitors have been reported, which may lead to undetected patient deterioration.

Sendelbach & Funk (2013) have demonstrated in their study that 72% to 99% of all alarms in critical care monitoring have no clinical relevance and result from measurement and movement artifacts. The medical staff may be exposed to approximately 187 audible alarms (Drew et al., 2014) and 700 physiologic monitor alarms (Cvach, 2012) per day for each patient.

#### **2.2.3.1. False Alarms - Causes**

Various factors can lead to FA including technical problems, artifacts, inappropriate alarm settings, alarm algorithm and so forth. Those factors can be divided into three categories (Imhoff & Kuhls, 2006):

- a) **Technical false alarms** – incorrect measurement of a parameter that appears to exceed a threshold and therefore triggers an alarm, although this parameter does not exceed any real predetermined threshold.
- b) **Clinical false alarms** – irrelevant (usually intermittent) clinical alarms even as the parameter is actually beyond the pre-determined alarm threshold. Caregivers do not adjust alarm thresholds for individual patients and tend to use default alarm settings due to the complex menu structure of the devices and a lack of training. The

general settings based on population parameters are more likely to produce FAs in critically ill patients (Drews, 2008).

- c) **False alarms through external interventions** – a combination of technical and clinical FAs caused by external interventions. Examples of these are moving the patient or disconnecting the patient from the ventilator for endotracheal suctioning.

In order to register the greatest percentage of clinically relevant events a high sensitivity in these alarm systems are essential. By design, most if not all of traditional threshold-based monitor algorithms are highly sensitive; however, this high sensitivity is achieved at the expense of specificity. As alarm limits become more sensitive and less specific, more FA are generated. This principle of alarm generation leads to an intolerable amount of alarms (Drew et al., 2004).

#### **2.2.3.2. False Alarm Outcomes**

The excessive quantity of alarms has led to a noisy alarm environment that constantly disturbs the patient's sleep. Alarms produce sound intensities above 80 dB that cause physiological stress for both patients and staff (Kam, Kam & Thompson, 1994). Furthermore, previous studies have shown that sudden alarms can cause an increase in heart and breathing rate, and depress the immune systems, thereby affecting recovery and length of stay in ICU (Xie, Kang & Mills, 2009). Willich et al. (2005) showed in their study that chronic noise increased risk of heart attacks by 50% for men and 75% for women.

Novaes et al. (1999) and Sorkin (1998) also emphasized the negative impact of the increased noise level in ICUs and found that machine alarms seem to disturb the members of the professional team even more than the patients themselves. Sometimes there may be so many different alarms at the same time that it makes it difficult for the clinicians to quickly identify the underlying condition. In some cases, the auditory signal is too loud and

shrill, inhibiting clinicians communicating at the time when communication is most necessary (Edworthy, 1994).

A repeated series of FA which eventually causes the medical staff to ignore the important alarms is called a 'cry- wolf' syndrome which leads to a lack of faith in the system and a casual attitude towards the constant presence of certain alarms (Sorkin, 1998).

Another phenomenon called alarm fatigue also occurs following high frequency of FA, commonly defined as desensitization (sensory overload) to alarm sounds, which causes caregivers to refrain from responding later to alarms, thereby raising serious patient safety concerns (Cvach, 2012). Caregivers may not pay attention to critical events and may falsely silence alarms or adjust the alarm setting beyond limits that are safe and appropriate for the patient (Sendelbach & Funk, 2013).

Moreover, the low nurse to patient (NTP) ratios and staffing levels discourage close vigilance, even when staffing is the most reliable factor in patient monitoring. It has been reported that only about 47% of all alarms are responded to by nurses (Gazarian, 2014), engendering sentinel events and patient deaths (Sendelbach & Funk, 2013).

### **2.2.3.3. Alarm hazard**

Excessive false and nuisance alarms may cause an alarm hazard, which includes dysfunction and disuse of alarm devices, alarm fatigue, and inappropriate alarm setting (DeVita et al., 2010).

In 2002, 65% of 23 sentinel events were related to inappropriate application of monitor devices and application of a uniform alarm range to every patient (Joint Commission, 2002). From 2005 through 2008, the Food and Drug Administration and the Manufacturer and User Facility Device Experience received 566 reports of patient deaths related to monitoring alarms in hospitals in the United States (Cvach, 2012). After the death of a patient



due to an alarm fatigue at Massachusetts General Hospital in 2010, the importance of alarm hazard was spurred (Wallis, 2010).

During the last decade, alarm hazards have been increasingly recognized as a major problem in the medical world. The Emergency Care Research Institute (ECRI) has ranked the clinical alarm hazard as the top priority for the fourth year in a row from 2012 to 2015 (ECRI Institute, 2011-2014), top 2 for 2016 (ECRI Institute, 2015) and top 3 for 2017 (ECRI Institute, 2016). This fact demonstrates very well the severity of the problems associated with alarms at ICUs.

#### **2.2.3.4. False alarms reduction**

The high rate of FA over the past two decades poses a significant yet unresolved concern in ICUs and should continue to be explored in further studies. This thesis aims at minimizing the number of false positive alarms and thus to increase patient safety in critical care. The high rates of FA as presented above suggest the potential for significant improvements by applying modern methods to identify and suppress FA in real time.

Multiple approaches have been identified to tackle the high frequency of FA in ICUs. Some investigations have focused on multi-sensors integration technique. Such a method was suggested by Aboukhalil et al. (2008), who demonstrated the potential of using multiple physiologic waveforms for reducing the incidence of false critical ECG arrhythmia alarms. By relating the ECG data with the arterial blood pressure curve their strategy reduced FA rate from 42.7% to 17.2%. Later Bitan & O'Connor's (2012), defined alarm conditions which replicate the logic of practitioners and correlate information across sensors. Their approach reduced FA frequency from 80% to 29%; but, their evaluation has been based on a limited database and missing data were not taken into account.

Recently Eerikäinen et al. (2016), developed an algorithm that selects the most reliable signal pair of ECG by comparing how well the detected beats match between different signals. Arrhythmia specific features are computed from the signal pair while the classification is performed with five separate random forest (RF) models. By their implementation false alarm rate (FAR) was reduced from 69% to 17%.

Some of the previous studies also attempted to suppress FAR through knowledge-based approaches enlisting expert knowledge. Koski et al. (1994) implemented a knowledge-based system for reducing FA on post-operative of cardiac surgery patients. In their test they achieved a sensitivity of 100% and the specificity increased from 20% to 73.9%; however, they didn't use a large, representative database for training or testing. Müller et al. (1997) proposed a knowledge-based alarm system for breathing circuit monitoring. The system introduced a TP rate of 93-100% and FAR that ranged from one false alarm per hour to one false alarm every 2.5 hours.

Some investigations have focused on learning methods based on expert knowledge. Tsien (2000) for example, used neural networks to detect events of interest in vital signs of pediatric ICU patients. The method has the drawbacks that the learning behavior is not reproducible, a long training phase is required, and training is not possible for primarily unstable patients.

Later Antink et al. (2016) proposed an approach combining multimodal rhythmicity estimation and several ML methods including linear discriminant analysis and RF. FAR was reduced to 22% while the results for some categories still need improvement.

In summary, different methods from diverse methodological fields have shown promising results in reducing FA noted in previous investigations; but, some were performed on small data sets or require a systematic validation procedure to evaluate the algorithm. Without such validation, it is

hard to believe that the algorithms will work well on unseen data. Additionally, none of these approaches has advanced into the mainstream of patient monitoring.

According to Imhoff & Kuhls (2006), methods that attempt to improve monitoring mechanisms must fulfill certain methodological criteria such as: (1) robustness against artifacts and missing values, (2) real-time application (i.e., efficient and fast algorithms), (3) predictable behavior and (4) methodological rigor.

The first criterion – missing data, are common in clinical studies and should be taken into account when modelling data for preventing interpretation biases (Little & Rubin 2014). Previous investigations into reducing FA in data including missing values are relatively few. Vesin et al. (2013) found that out of 44 published clinical studies, 36% did not make any mention of missing data. Worse still, less than 5% studies acknowledged the importance of missing data and the need to address the problem.

The approach used in this work tries to imitate the way medical staff assess a patient's condition by correlating several parameters at once and considering their relations among each other. This method has presented to significantly reduce the number of FA in previous studies. The hypothesizing of this thesis is that applying an ML data-driven approach toward developing a clinical alarm model that overcomes missing data in sensors of medical monitors will help reduce FAR reported in ICUs in situations of missing parameters.

Unlike previous studies, This study suggests a practical approach that focuses on depressing FAR by dealing with missing data. Previous chapters reviewed STD tools and the background of the problem this study faces. The next section will review ML methods by which this study was able to deal with missing values and reduce the FAR.

## **2.3. Decision Trees and Forests**

### **2.3.1. Machine Learning overview**

Data analysis and ML have become an integral part of the modern scientific methodology, offering automated procedures for the prediction of a phenomenon based on past observations, unraveling underlying patterns in data and providing insights about the problem (Samuel, 1959).

Classification is a supervised learning task that aims at teaching a system to make decisions or perform predictions based on labeled observations. The algorithm analyzes the training data and produces an inference that enables the system to automatically assign a new observation to one of the previously defined classes. The chosen classification method in this research is RF, algorithm that will be review in this chapter (Mohri et al., 2012).

### **2.3.2. Decision Trees**

Decision trees are predictive models that are used for regression as well as classification problems. Regression trees have a continuous response variable and classification trees have a discrete response variable, while in this thesis classification trees is the area of interest. A tree is made up of a set of connected nodes and edges arranged in a tree-like structure, while each node represents a decision. The decision is based on a split in one of the predictors. Each classification tree is built of a metric that relies on random sampling with a replacement called bootstrapping. This statistic method creates new datasets from existing data and the observations are picked randomly with equal probability for all observations.

The terminal nodes are the leaves of the tree that store the class labels. Trees are built to optimize a certain function, which involves selecting features to create the best split on the internal node. A classification tree partitions the problem recursively into sub problems, treat these regions as separate data sets and

employ the same tactic of finding a good split for the top node. It will continue down recursively until all data have been separated in leaves. There are several algorithms for choosing the decision in the root node such as gain ratio, information gain, and Gini index (Quinlan, 1986).

#### **2.3.2.1. ID3**

Iterative Dichotomiser 3 (ID3) is one of the first decision-tree construction algorithms, invented by Ross Quinlan in 1986. Each iteration of ID3, contains a previously unused attribute associated with the biggest information gain. This attribute is selected to split the set of observations at a node, beginning at the root of the tree. Recursion on subsets of internal nodes using remaining attributes may stop in one of these cases:

- a) Every item of a node belongs to the same label, then the node becomes a leaf and labeled with the class of the examples.
- b) There are no more attributes to be selected, in which case the node turned into a leaf labeled with the most common class of the examples in the subset.
- c) A situation in which a subset after a split is empty. In this case the parent node becomes a leaf labeled with the class of the most frequently occurring element in the node.

During testing, each observation is classified by the traversing down the decision tree from the root to a leaf node, which contains the predicted label. At runtime, the tree is used to classify new unseen objects by traversing down the decision tree from the root to a leaf node that contains the predicted class.

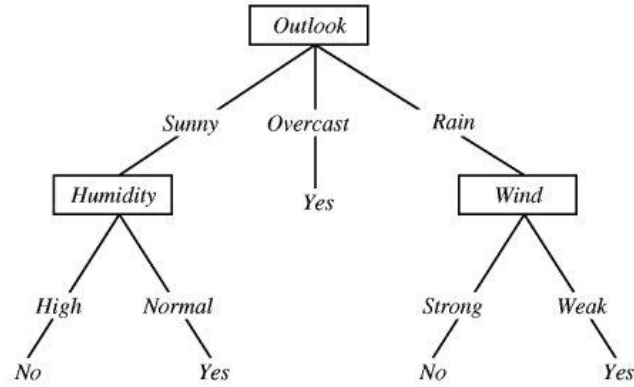


Figure 3 - A decision tree for concept Play Tennis, an example is classified by sorting it through the tree to the appropriate leaf node, then returning the classification associated with leaf (in this case Yes or No). This tree classifies Saturday morning (Mitchell, 1997).

ID3 growth occurs in each branch of a tree just deeply enough to perfectly classify the training set which may lead to overfitting. The reasons for overfitting are a lack of representative instances if data size is small or noise in the data. For avoiding overfitting, there are two main approaches, the first one is to stop growing the tree earlier, before it perfectly classifies the training set, and the second is to allow overfitting but then post-prune the tree.

#### 2.3.2.2. C4.5

C4.5 algorithm was invented by Quinlan (1993), as an improvement to ID3. C4.5 can handle both continuous and discrete attribute values, as well as a training set with missing attribute values. C4.5 are simply do not use missing attribute values in gain and entropy calculation during the process of choosing the attribute that best satisfies the splitting criterion. In addition, C4.5 decision trees goes back through the tree once it has been created and attempts to remove irrelevant branches by replacing them with leaf nodes. This pruning process improves accuracy by reducing overfitting.

#### 2.3.2.3. CART

The Classification and Regression Trees (CART) method was first introduced by Breiman et al. (1984). CART trees is similar to C4.5

algorithm and can be optimized to perform classification or regression. CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample.

### 2.3.3. Attribute Selection Measures

Most of the decision tree algorithms adopt a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Although those algorithms differ in many aspects, the main differences are in their quantities measure to attribute discriminatory ability and pruning tree methods. The next sections will present some attribute selection measures and pruning tree methods that are commonly used.

#### 2.3.3.1. Information Gain

Information gain is a statistical property that helps to decide which attribute discriminates the training data according to target classification most accurately and is usually used for multiclass classification problems. For selecting the attribute that is most useful for classifying the examples, the entropy characterizes impurity of a set of examples and the attribute with the biggest decrease in entropy if being chosen as a split point.

The entropy of a data set is given by

$$Entropy(D) = -\sum_{i=1}^m p_i \log_2 p_i \quad (4)$$

Where  $p_i$  is the probability that an instance in set  $D$  belongs to class  $C_i$ . It is calculated by  $|C_i|/|D|$ . Suppose the attribute  $A$  is now considered to be the split point and  $A$  has  $v$  distinct values  $\{a_1, a_2, \dots, a_v\}$ . Attribute  $A$  can be used to split  $D$  into  $v$  subsets  $\{D_1, D_2, \dots, D_v\}$  where  $D_i$  consists of instances in  $D$  that have outcome  $a_j$ . The new entropy is defined by the following equation.

$$Entropy_A(D) = -\sum_{j=1}^v \frac{|D_j|}{|D|} \times Entropy(D_j) \quad (5)$$

The information gain when using attribute  $A$  as a split point is as follows:

$$Gain(A) = Entropy(D) - Entropy_A(D) \quad (6)$$

$Gain(A)$  presents how much would be gained by branching on  $A$ . Therefore, the attribute  $A$  with the highest  $Gain(A)$  should be chosen to use (Mitchell, 1997).

### 2.3.3.2. Gain Ratio

The information gain measure presented above is bias toward attributes having a large number of values, thus leading to a bias toward tests with many outcomes. C4.5, a successor of ID3, uses an extension to information gain called Gain ratio, which attempts to overcome this shortcoming. The method normalizes information gain by using a split information factor, defined as follows.

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left( \frac{|D_j|}{|D|} \right) \quad (7)$$

Gain ratio is then given by the following equation.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (8)$$

The attribute with the highest gain ratio is selected as the splitting point (Mitchell, 1997).

### 2.3.3.3. Gini Index

The CART algorithm uses the Gini index as its attribute selection measure. The Gini index measures the impurity of a set; therefore, it is also called Gini impurity. The impurity of set  $D$  by Gini index is defined as follows (The notation is the same as in the previous methods):

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2 \quad (9)$$

Suppose the attribute  $A$  is now considered to be the split point and  $A$  has two distinct values  $a_1, a_2$ . Attribute  $A$  can then be used to split  $D$  into



$D_1$  and  $D_2$  where  $D_i$  consists of instances in  $D$  that have outcome  $a_j$ . The Gini index only considers a binary split for each attribute as the following equation.

$$Gain_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (10)$$

The reduction in impurity is defined as:

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \quad (11)$$

The attribute with the highest reduction in impurity is selected for the next classification step (Breiman et al., 1984).

### 2.3.4. Random Forests

Decision trees are easy to understand and to implement irrespective of the size of the dataset and can handle variability in the attribute value types (numerical, categorical). However, trees constructed with greedy algorithms may not yield a globally-optimal solution. One problem associated with individual decision trees is that they tend to over-fit and describe random error or noise instead of the underlying relationship. This problem leads to poor predictive performance on the test set.

In the family of ensemble learning methods, RF proposed by Ho in 1995, is regarded as one of the most powerful ones. Consisting in an ensemble of independent decision trees, RFs are very intuitive models that offer a flexible probabilistic framework for solving different learning tasks. It exploits the power of many decision trees, judicious randomization to generate accurate predictive models. Moreover, it also provides insights into parameters importance, missing value imputations, etc. The RF has remarkably few controls to learn, and therefore, analysts can effortlessly obtain effective models with almost no data preparation or modeling expertise. Besides, short training time and the ability to run in parallel are two other huge advantages of the RF.

For dealing with the problem of overfitting, an element of randomness is used in these ensembles of decision trees, while each tree is built from an independently and randomly sampled subset of available observations. RF is useful for prediction problems and supervised-learning tasks in various fields, including biological science, finance, chemical engineering, agro-science, medical analysis, etc. This ensemble method is a combination of multiple classification or regression trees, which engages in a voting strategy to provide the final prediction. The trees are combined into a larger RF model where each tree casts a vote on the predicted class, and the predicted label corresponds to the class with the most number of votes. This research focuses on classification problems most relevant to FAR reduction.

Previous studies have shown the impressive predictive performance of RF in various fields (Siroky 2009; Kumar & Thenmozhi 2006; Shi et al. 2005; Ward et al. 2006; Diaz-Uriarte & De Andres 2006; Jiang et al. 2007; Goldstein et al. 2011). The main principle behind ensemble methods is that a group of “weak learners” can come together to form a “strong learner”. At each node of a tree, a fixed-size random subset of features is chosen out of all available features, and the feature on this subset that best separates the data is selected to split the node. During classification, a new test object is classified by multiple decision trees in the forest and outputting the class label that receives the maximum number of votes.

RF models perform well in many classification tasks and work efficiently on large datasets. Each tree in the forest is built using all the training samples sampled with replacement. While training a tree, each node has access to only a randomly chosen subset of the entire set of parameters and is trained to optimize the parameters in each node of every tree. RFs have been shown to be robust to the effects of noise and outliers. Moreover, they generalize well to variations in data. Thus, RFs are suitable for classification tasks involving medical data such as patient's monitoring collected by sensors which can be sensitive to noise and can exhibit a high level of variance.

## 2.4. Tools for Diagnostic Tests Performance Evaluation

The obtained results of classification problems such as systems that involve detection, diagnostics, or prediction had to be assessed to validate the discriminative power of a given analysis. However, depending solely on the quantization of Hits and Misses of a test group is not enough for system quality evaluation, since it depends on the quality and distribution of the test group data.

In recent decades, researchers have developed numerous statistical methods to evaluate the performance of binary classifiers. Common measures in literature for evaluating results of ML experiments are Recall, Precision and F-measure. These measures are generally used in areas where one is primarily interested in finding the "positives" such as Performance Marketing or the area of Information Retrieval (Manning & Schutze, 1999; Raghavan et al., 1989). The drawback of these measures is the disregard for the "true negative" count; thus, these measurements can be misleading and present some biases.

Accuracy was also an acceptable measurement; but was also found as a poor metric for measuring performance. The higher severity of false negative errors (over false positive errors) is not reflected in the naïve accuracy measure or another statistical measures normally used (Provost, Fawcett & Kohavi, 1998).

Current research has shifted away from simply presenting the results of these measures when performing an empirical validation of new algorithms in the ML domain. This holds especially true when evaluating algorithms that output probabilities of class values. The output of typical classification models, such as RF, is a binary label and a measure of confidence in the prediction. The next chapter will review some alternate techniques to assess binary classifiers using this measure of confidence that overcomes the bias problem.

### **2.4.1. ROC curve as an Assessment Measurement in Machine Learning**

In the Medical Sciences, ROC analysis has been borrowed from SDT (presented in previous chapter) to become a standard for evaluation and standard setting, comparing TP rate and FAR (Green & Swets, 1966).

Spackman (1989) was the first who demonstrated the value of ROC curves in comparing and evaluating different classification algorithms of ML. In this domain, ROC curve presents how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. It illustrates the performance of a binary classifier system as its discrimination threshold is varied. ROC curves presents a visualization of the trade-off between Hit rates and FAR of classifiers.

### **2.4.2. AUC**

Regarding medical diagnosis and a ML community, one of the methods for combining these measures into the evaluation task is the analysis of the area under the ROC curve (AUC). The AUC is used as a simple metric to define how an algorithm performs over the whole space (Singla & Domingos, 2005) which is equivalent to the Wilcoxon test of ranks (Hanley & McNeil, 1982). It is helpful to compare classifiers and create a two-dimensional depiction by reducing their ROC performance to a single scalar value. In practice, AUC performs very well and is often used when a general measure of predictiveness is desired.

Since AUC is a portion of the area of the unit square, its value will always be between 0 and 1. However, because random guessing produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a

randomly chosen positive instance higher than a randomly chosen negative instance (Bradley, 1997; Hanley & McNeil, 1982).

### 2.4.3. Confusion matrix

In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table (Stehman, 1997). This is a technique for summarizing the performance of a classification algorithm. Terms such as accuracy, sensitivity, specificity, PPV, Cohen's Kappa, F Score and ROC Curve can be computed through this matrix and performance of such algorithms is commonly evaluated using those.

Confusion matrix has four categories: True positives (TP) are examples correctly labeled as positives. False positives (FP) refer to negative examples incorrectly labeled as positive and is also known as Type I error. True negatives (TN) correspond to negatives correctly labeled as negative. Finally, false negatives (FN) refer to positive examples incorrectly labeled as negative and is also known as Type II error.

Previous chapter has used the above terms as following:

- TP = Hits
- FP = FA
- TN = Correct Rejections
- FN = Miss

The structure of confusion matrix for a binary classifier is composed as follows:

| <b>Predicted \ Actual</b> | <b>Positive</b> | <b>Negative</b> |
|---------------------------|-----------------|-----------------|
| <b>Positive</b>           | TP              | FN              |
| <b>Negative</b>           | FP              | TN              |

Table 2 - confusion matrix for a binary classifier.

#### 2.4.4. Youden's index

The index was suggested by W.J. Youden in 1950 as a way of summarizing the performance of a diagnostic test. Youden's index ( $J$ ) is a statistic combines sensitivity and specificity into a single measure and is also used in SDT as the discriminability index ( $d'$ ). This measurement ranges between 0 and 1, while, in a perfect test, will be equals 1. Although the theoretical range of the Youden's Index is from -1 to 1, the practical range in use is often from 0 to 1 since negative values of the Youden's Index do not have meaningful interpretation in practice.  $J$  is represented by the vertical distance between the ROC curve and the first bisector (or chance line).

Maximizing this index allows us to find, from the ROC curve, an optimal cut-off point ( $c^*$ ) because it is optimizes differentiating ability when equal weight is given to sensitivity and specificity.

$J$  can be formally defined as (Božikov & Zaletel-Kragelj, 2010):

$$J = \max_c \{ \text{Sensitivity}(c) + \text{Specificity}(c) - 1 \} \quad (12)$$

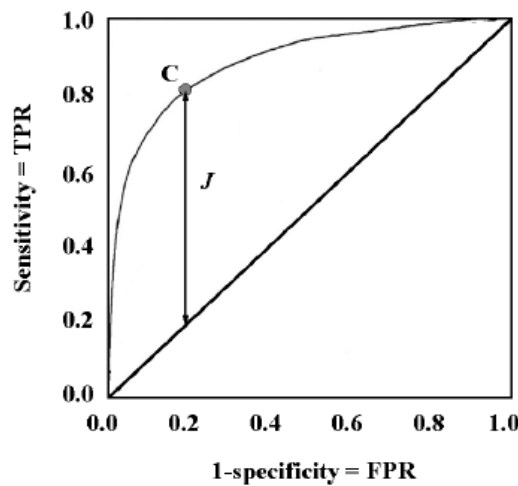


Figure 4 - Finding best cut-off from the ROC curve by using Youden's index  $J$  (Božikov & Zaletel-Kragelj, 2010).

False negatives and false positives are significant issues in medical testing; missing any problem in patient's condition might be critical, but also an excessive rate of FA can lead to alarm fatigue as presented in previous chapters.

In order to evaluate the quality of the classifiers and to find the optimal balance between sensitivity and specificity the above statistical measures were used, which display the trade-off between these terms and is useful in assigning the best cut-offs for clinical use.

## **2.5. Literature Review Summary**

The literature review provided in this section cover a few topics and issues that are relevant to challenges this research attempts to address – FA reduction in ICU Monitors. Although some significant research tried to find solution to high FA rates, this problem has been studied extensively in the last two decades with no proper solution found. As reviewed in literature, ECRI recently listed its 2017 Top 10 Health Technology Hazards, the alarm hazard placed among the top 3 for the sixth year in a row. This fact demonstrates very well the severity of the problems associated with alarms at hospitals and the need to continue to explore this issue.

The goal of this thesis is to demonstrate one approach to suppressing FAR in situations of missing data using ML methodologies based on expert knowledge. The main hypothesis of this work is that training ML algorithm to integrate information from the available sensors will compensate for the missingness and might decrease the FARs.



### **3. Research Methodology**

This study was conducted under the supervision of the Helsinki Committee of Tel-Aviv Medical Center (TAMC), and was approved by the Ethics Committee of Ben-Gurion University.

As illustrated in the literature review, the necessity of an accurate methodology for FA reduction in ICUs is well acknowledged. This thesis applies expert-based rules that enable real time data analysis and decision making for patient monitoring. This research has attempted to optimize this approach in cases of missing parameters using a RF framework. First, the chapter will introduce the development of the algorithm, then demonstrate the clinical data that were used in this research, and finally present the result evaluation.

#### **3.1. Model Overview**

Bitan & O'Connor (2012) demonstrated an approach for suppressing FAR using expert-based rules, which correlates information across sensors and tries to validate alarms by imitating specialist reasoning. They found that matching different parameters can be used for the reduction of FAR, and can employ more sophisticated rules of alarm scenarios than single-sensor based approaches. However, physiological data were collected by separate sensors in unsynchronized times, which led to missing data and resulted in poor performance of the developed algorithm. Furthermore, the inability of the algorithm to deal with missing sensor data caused an increase in FAR, which makes the algorithm less effective.

This research attempts to address the missing sensor data issue using ML, which can consolidate information from available sensors to compensate for missing sensors, and thereby extend Bitan & O'Connor (2012). RF models were trained based on clinical parameters to mimic decisions based on the FER. The classification made by the FER model, which is equivalent to that of the medical

specialist, was used as a ground truth in model training and performance evaluation. Then RF models were trained based on partial sets of parameters, i.e., without one, two, or three of the parameters. In these partial settings, the FER model classification was used as a ground truth. The classification test results of the RF models trained using the partial sets of parameters and were compared with those of the compatible PER to check which of the two – the medical specialist or the ML-based model, both relying on missing data – is more accurate.

The partial models are stored in a "bank" of alarm classifiers that can be used for detecting alarms in ICU monitoring systems, even when the measuring of several parameters has failed. For example, if a 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 that has been trained on data that does not include the parameter will be extracted from the bank. This pre-trained model, which does not require ARTBPS as an input parameter, will be used in the monitoring system to assist the medical staff in assessing the patient's condition.

### **3.2. Data resource**

The data in this research was retrieved from samples of anonymous patients who were admitted to the Tel-Aviv Medical Center (TAMC) ICU between 2008 and 2014. Data were derived using the Metavision software, which tracked information in the 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 from F-CUB 08 model and have a high sampling frequency (hundreds of samples per second); but, the data is deleted each month and isn't stored in the system.

The monitor alerts caretakers when it has a signal problem and the values exceed a predefined default threshold. The medical crew adjusts the alarm thresholds for individual patients but tends to use the default alarm settings. The general

settings based on population parameters are more likely to produce FA in critically ill patients (Drews, 2008). These values do not change for 85-90% of the patients. In the remaining cases the nurses, with the approval of a senior physician, change the threshold values. The data stored by the Metavision software isn't recorded by the monitoring equipment; thus, it doesn't include the monitor alarms. To overcome this problem monitor alarms have been added according to the default alarm settings; however, TAMC doesn't contain the changes on settings that were made by the caregivers that could decrease the FAR. Additionally, the alarm information about the clinical function or an external disturbance (e.g., someone moved the bed) is not included.

The database contains 681,265,089 samples obtained from 7,688 patients. All the monitor alarms at the database were annotated by default alarm settings as following:

| Parameter Name                 | Limits      |
|--------------------------------|-------------|
| Heart Rate                     | 60-120 bpm  |
| Non Invasive Arterial Pressure | 90-180 mmHg |
| Arterial Pressure Systolic     | 90-180 mmHg |
| Mean Arterial Pressure         | 65-125 mmHg |
| Saturation                     | 90-100 %    |
| Central Venous Pressure        | 0-15 mmHg   |

*Table 3 - Default alarm settings.*

The database is designed using SQL server desktop version. Different patients were monitored for different sets of physiological parameters. Some parameters were measured for all patients while others were measured only for few. The elaboration on the choice of parameters for the model presented in this study, in the following section.

### **3.3. Data understanding**

#### **3.3.1. Independent Variables - Medical Monitoring Parameters**

This section will briefly describe each of the parameters used in this research.

- **Heart rate (HR)**

HR is the number of contractions of the heart in one minute, expressed in beats per minute (bpm).

- **Arterial Blood Pressure Systolic (ARTBPS)**

ARTBPS refers to the maximum pressure exerted on blood vessels in systemic circulation because of the contraction of the left ventricle of the heart.

- **Mean Arterial Blood Pressure (ARTBPM)**

ARTBPM calculates the average arterial pressure in a patient's arteries during from measured systolic (maximum) and diastolic (minimum) blood pressure values.

- **Central Venous Pressure (CVP)**

CVP is the direct measurement of the blood pressure in the central veins near the right atrium of the heart.

- **Pulmonary Artery Pressure Diastolic (PAPD)**

PAPD is lowest pressure generated by the right ventricle ejecting blood into the pulmonary circulation. In patients with a normal pulmonary vascular resistance, it correlates with the Left Ventricular End Diastolic Blood pressure, and can be used as a proxy for its value.

- **Total Respiratory Rate (RR total)**

RR total also known as the breathing rate means the number of breathing cycles in one minute, measured in respirations per minute (rpm). In spontaneously breathing patients, it is the same as the “respiratory rate”. In

mechanically ventilated patients, it is the sum of breaths initiated by the ventilator and any efforts above that made by the patient. Respiration rates may increase with medical problem as illness, fever etc.

- **Mechanical Respiratory Rate (RR mandatory)**

RR mandatory is a device that supports or completely controls breathing depending on the patient's condition. The mandatory rate is the rate of breathes that the machine will deliver to a patient who is not breathing, or is not making any efforts above this rate.

- **Saturation Peripheral Capillary Oxygen (Spo2)**

Spo2 is defined as the percentage of oxygenated hemoglobin (hemoglobin containing oxygen) compared to the total amount of hemoglobin in the blood (oxygenated and non-oxygenated hemoglobin). It is measured at the bedside using a pulse-oximeter, which also measures and reports the heart rate.

- **Fraction of inspired oxygen (Fio2)**

Fio2 is defined as the percentage of oxygen concentration that being received by the patient and can be set on modern mechanical ventilators, or estimated when the patient is being treated with supplemental oxygen via a variety of appliances (e.g. nasal cannulae, face mask, etc).

- **ST Segment**

The ST segment is the flat, isoelectric section of the ECG between the end of the S wave and the beginning of the T wave. It represents the interval between ventricular depolarization and repolarization. It can be elevated or depressed by a variety of conditions, the most common of which is myocardial ischemia.

The table below (Table 4) presents the normal ranges of the above parameters dividing by categories:

| Category              | Parameter Name | Description                                 | Normal Range             |
|-----------------------|----------------|---|--------------------------|
| <b>Blood Pressure</b> | ARTBPS         | Arterial Pressure Systolic                  | 90 - 140 mmHg            |
|                       | ARTBPM         | Arterial Pressure Mean                      | 70 - 105 mmHg            |
|                       | CVP            | Central Venous Pressure                     | 3–8 mmHg                 |
|                       | PAPD           | Pulmonary Artery Pressure Diastolic         | 8 - 15 mmHg              |
| <b>Heart Beat</b>     | HR             | Heart Rate                                  | 60–100 bpm               |
|                       | ST Segment     | Section of the ECG between the S and T wave | -0.5 mm< and < 0.5 mm    |
| <b>Respiration</b>    | RR Mandatory   | Mechanical Respiratory Rate                 | 12-18-breaths per minute |
|                       | RR Total       | Total Respiratory Rate                      | 12-18-breaths per minute |
|                       | Spo2           | Oxygen Saturation                           | > 92%                    |
|                       | Fio2           | Fraction of Inspired Oxygen                 | 30%-50%                  |

Table 4 - Normal ranges of parameters for an adult.

### 3.3.2. Dependent variable

"FER Alarm" parameter is the "Alarm" or "No Alarm" classification that was determined by the developed FER per minute of sampling.

| Parameter Name   | Description    | Type   |
|------------------|----------------|--------|
| <b>FER Alarm</b> | 1 – True Alarm | Binary |
|                  | 0 – No Alarm   |        |

Table 5 - Explained variable.

### 3.3.3. Clinical Alarm Scenarios

This section will briefly describe each of the clinical alarm scenarios that were examined in this research.

- Hypotension**

Hypotension is low blood pressure as measured with a blood pressure cuff or arterial line. Severely low blood pressure can deprive the brain and other

vital organs of oxygen and nutrients, leading to a life-threatening condition called shock.

- **Hypovolemia**

Hypovolemia or volume contraction is sometimes used synonymously and demonstrates a condition of decreased blood volume. Common causes of hypovolemia are loss of blood (external or internal bleeding or blood donation), loss of plasma (severe burns and lesions discharging fluid), loss of body sodium and consequent intravascular water; e.g. diarrhea or vomiting.

- **Left ventricular (LV) shock**

LV shock is a life-threatening medical condition that is characterized by pulmonary edema and low output-hypotension attributable to left ventricular systolic and diastolic dysfunction resulting from acute ischemia-infarction often superimposed on prior infarction.

- **Obstructive shock**

Obstructive shock is a form of shock associated with physical obstruction of the great vessels or the heart itself. Tension Pneumothorax, auto-PEEP, abdominal compartment syndrome, and cardiac tamponade are causes of obstructive shock.

- **Bradycardia**

Bradycardia is a condition of an abnormally slow heart rate. The heart is not able to pump enough oxygen-rich blood to your body. Bradycardia symptoms include fatigue, weakness, dizziness, sweating, and at very low rates, fainting/loss of consciousness. Patients with severe bradycardia (e.g. HR, 30/min) will have a cardiac output inadequate to perfuse their vital organs, and will die.

- **Tachycardia**

Tachycardia is a common type of heart rhythm disorder (arrhythmia) in which the heart beats faster than normal. This state can originate from either

the upper heart chambers (atrial tachycardia) or lower heart chambers (ventricular tachycardia).

(Marino, 2013)

### 3.3.4. Description statistics

#### 3.3.4.1. Data Visualization Tool

In order to understand the abnormal behaviors and enable patterns to be analyzed conveniently, an interactive data visualization tool was built. This tool allows easy access to the TAMC database and classified patient clinical data. It also provides flexibility in selecting parameters from categorized lists that can be visualized, and present advanced visualization dimensions. These visualizations help to detect patterns using drill down abilities on the recorded parameters. The platform was designed by JavaScript for building the frontend and PHP for the backend. The source code of this tool can be found on GitHub repository at:

<https://github.com/galhev/Thesis/tree/master/VisualizationTool>.

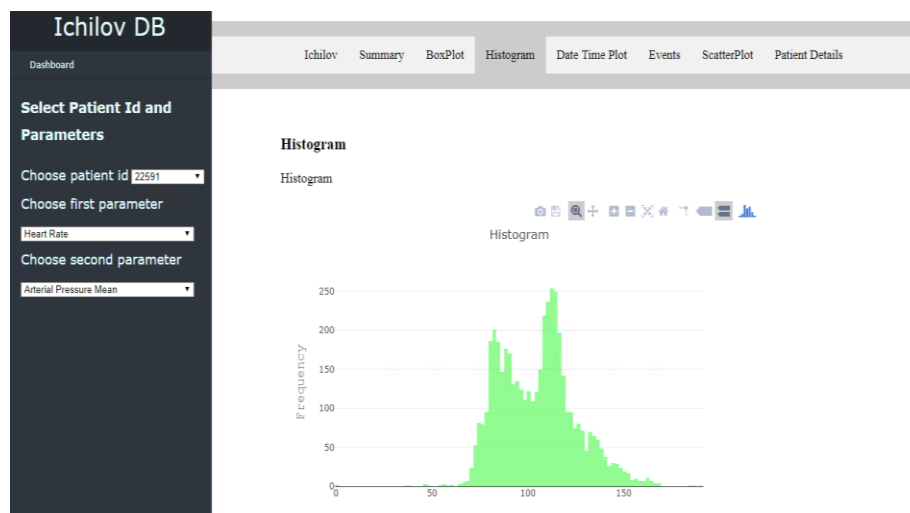


Figure 5 - Histogram Tab in the data visualization tool.

The classification feature enables a convenient way to mark a specific point in time as an abnormal by clicking and coloring in red. By hovering over a



particular point, it is possible to see other parameters that were sampled in the same minute.

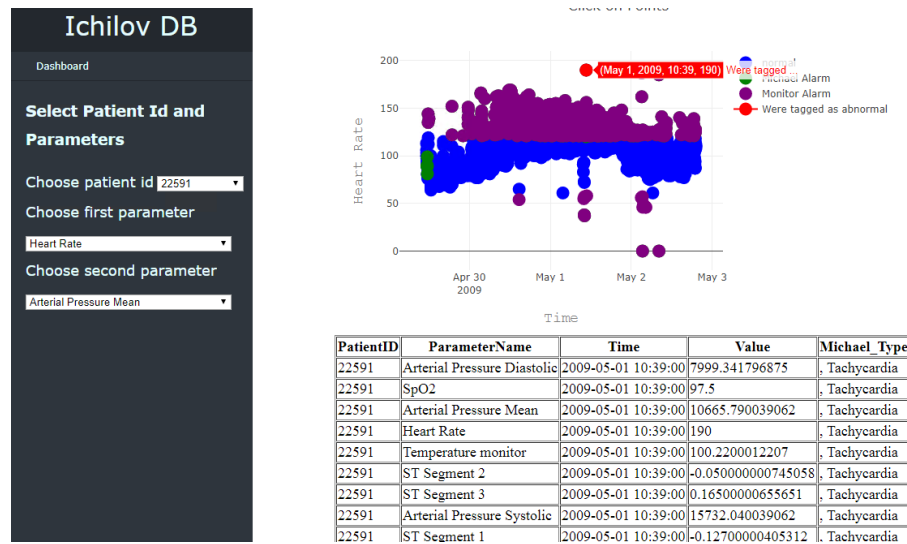


Figure 6 - Classification feature in the visualization tool.

Descriptive statistics about physiological parameters are attached in Appendix B.

### 3.3.4.2. Alarm Frequency

As part of the data analysis process, the prevalence of clinical and technical alarms was examined in order to explore any particular period of time when these alarms were more frequent. The total number of these alarms per bed per hour was plotted against time, as shown in Figures 8.

The duration of shifts occurs as following:

**Morning:** 07:00 to 15:00

**Afternoon:** 15:00 to 23:00

**Night:** 23:00 to 07:00

On average, there are 12-16 patients per day in ICU. Morning and afternoon shifts have higher doctors' staffing levels, with a ratio of approximately 1:3, while during the night shift the maximum ratio is 1:6 or even less. The nursing staff level is equal throughout the day, with a NTP ratio of approximately 1:2.

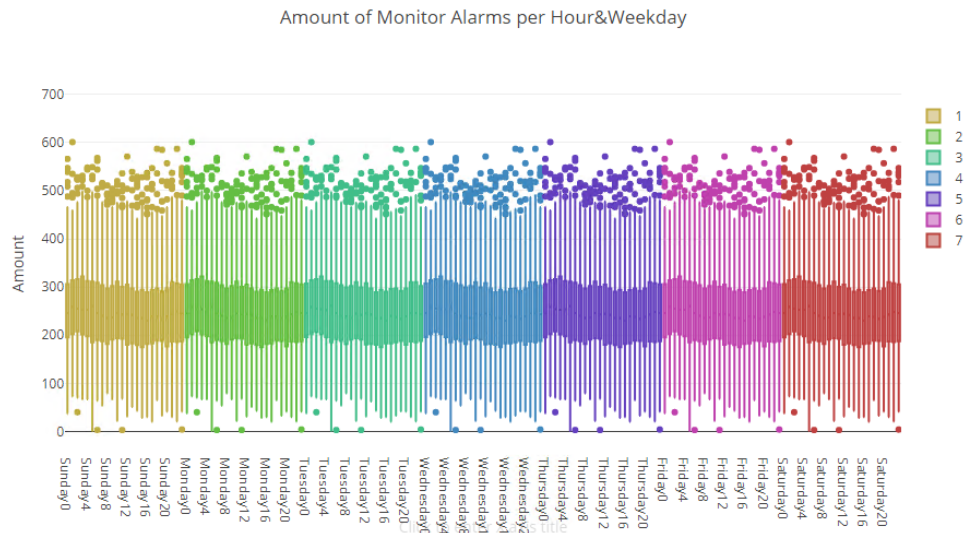


Figure 7 - Total daily monitor alarm (clinical and technical) distribution among day and hour.

The prevalence of alarm frequency by hour and weekday available at this link: [Amount of monitor alarms per hour and weekday.](#)

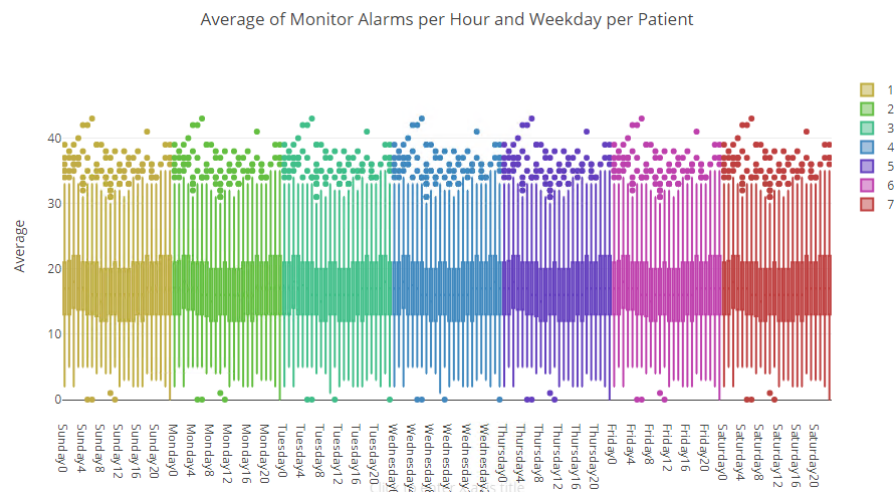


Figure 8 - Average per patient of daily monitor alarm (clinical and technical) distribution among day and hour.

The prevalence of alarm frequency by hour and weekday per patient available at this link: [Average of monitor alarms per hour and weekday.](#)

| Shift     | Average | Median | SD    | MAX | MIN |
|-----------|---------|--------|-------|-----|-----|
| Morning   | 250.99  | 245    | 87.34 | 695 | 1   |
| Afternoon | 245.40  | 239    | 87.97 | 635 | 1   |
| Night     | 262.18  | 256    | 89.21 | 656 | 1   |

Table 6 - Monitor alarms per shift.

| Hour | Average  | Median | SD       | MAX | MIN |
|------|----------|--------|----------|-----|-----|
| 0    | 249      | 249    | 90.2573  | 570 | 2   |
| 1    | 257      | 257    | 87.29967 | 601 | 6   |
| 2    | 258      | 258    | 87.27913 | 654 | 3   |
| 3    | 256.5    | 256.5  | 88.99508 | 656 | 31  |
| 4    | 253      | 253    | 89.51987 | 652 | 3   |
| 5    | 258      | 258    | 90.96712 | 625 | 1   |
| 6    | 253      | 253    | 88.47807 | 568 | 3   |
| 7    | 251      | 251    | 85.70576 | 594 | 34  |
| 8    | 240      | 240    | 87.03198 | 553 | 13  |
| 9    | 244.5    | 244.5  | 85.8848  | 544 | 65  |
| 10   | 249      | 240    | 87.03198 | 553 | 13  |
| 11   | 244.5    | 244.5  | 88.00089 | 545 | 3   |
| 12   | 248.5014 | 244.5  | 89.85136 | 571 | 31  |
| 13   | 242.9589 | 240    | 87.03198 | 553 | 13  |
| 14   | 239.6507 | 235.5  | 88.25821 | 529 | 43  |
| 15   | 240.1164 | 232    | 88.40978 | 528 | 28  |
| 16   | 237.3726 | 228.5  | 86.65689 | 537 | 24  |
| 17   | 239.5575 | 229    | 85.70449 | 547 | 20  |
| 18   | 243.4411 | 236    | 88.06903 | 589 | 39  |
| 19   | 236.7274 | 227.5  | 86.85984 | 585 | 14  |
| 20   | 240.6479 | 232    | 89.36624 | 635 | 41  |
| 21   | 248.2973 | 237    | 88.75077 | 555 | 30  |
| 22   | 252.9575 | 245    | 90.00723 | 587 | 21  |
| 23   | 252.5575 | 243    | 90.96262 | 547 | 4   |

Table 7 - Monitor alarms per hour.

Figure 7 and Figure 8 clearly show that the alarm distribution among the day is cyclic. Moreover, the alarm frequency varies throughout the day. As presented in Table 6, the amount of alarms in the night shifts is higher than afternoon and morning while the lowest alarm frequency occurs during afternoon. As shown on Table 7, the high frequency monitor alarms occur between 01:00 to 05:00 am and the low frequency occur between 16:00 am to 19:00 pm. For the whole description, see [Appendix C](#).

### 3.3.4.3. Sampling Frequency and Missing Data

Table 8 lists the twelve parameters chosen for the study with their frequency in the TAMC database and percentage of missingness. Percentage of missingness was calculated by the ratio between the empty samples (of patients for whom the parameter was sampled at least once) and the total amount of samples of those patients.

|   | <b>Parameter Name</b> | <b>% of Patients</b> | <b>% of Missing Data</b> |
|---|-----------------------|----------------------|--------------------------|
| <b>Always Monitored parameters</b>          | HR                    | 99.83                | 1.7                      |
|   | Spo2                  | 99.66                | 9.17                     |
|   | ST1                   | 97.06                | 8.69                     |
|   | RR Total              | 94.66                | 56.58                    |
| <b>Frequently monitored parameters</b>      | ST2                   | 88.85                | 21.9                     |
|   | ST3                   | 88.85                | 21.9                     |
|   | Fio2                  | 88.18                | 60.31                    |
|   | ARTBPS                | 80.06                | 17.77                    |
|   | ARTBPM                | 80.39                | 17.8                     |
|   | RR Mandatory          | 79.76                | 56.12                    |
| <b>Less frequently monitored parameters</b> | CVP                   | 24.88                | 64.98                    |
|   | PAPD                  | 3.91                 | 85.06                    |

*Table 8 - Parameter frequency and missingness in the TAMC database.*

As reported in other clinical studies, missing data are common and unavoidable. Their existence also presented in TAMC clinical database. The next sections will show the impact of missing values on the performances of the method suggested by Bitan & O'Connor's (2012) and offer the solution of this research to handle with this problem.

### **3.4. Data Preparation**

TAMC database contains two main and most important tables for this research – "Signals" and "Parameters". "Signals" table contains a list of parameters and their values for each patient while "Parameters" table contains details of a list of 4,534 physiological parameters. In order to design the raw database to be convenient and easy to work with, a dataset of 7,267 patients was extracted and reshaped. Each sample in the extracted dataset portrays simultaneous measuring of twelve physiological parameters for a single patient at a specific minute. The extracted parameters are: HR, ARTBPS, ARTBPM, CVP, PAPD, RR Total, RR Mandatory, Spo2, Fio2, ST1, ST2, ST3. Some of the parameters were chosen because of their medical importance (as part of the developed FER) as HR, ARTBPS, ARTBPM, CVP, PAPD, RR Total, RR Mandatory. The other parameters were selected because of their high appearance prevalence in the database.

### **3.5. Expert-Based Rules Implementation**

This research applies expert-based rules to tag clinical alarms and diagnoses a patient's medical state by a data-driven model that is trained to map multiple signals across monitor sensors onto the tagged alarms. According to the suggested approach, an alarm is triggered when one or more parameters cross their threshold values at a predetermined time. Following, each sample in the dataset was tagged as "Alarm" or "No alarm" according to a set of classification rules defined by Prof. Michael F. O'Connor based on the measured values of five parameters: HR, ARTBPS, ARTBPM, CVP and PAPD. Prof. Michael F. O'Connor, M.D. serves as the Director of Critical Care Medicine at the University of Chicago Medical Center. As an anesthesiologist and intensivist, Dr. O'Connor spent much of the past 23 years working in intensive care units, including the Burn ICU, Surgical ICU, Medical ICU and Cardiothoracic ICU. His wide range of clinical interests includes improving the performance of alarms in the ICU and predicting decompensation in ward patients.

Prof. O'Connor considered alarms triggered in seven clinical scenarios: Bradycardia, Bradycardia hypotension, Hypovolemia, Tachycardia, Tachycardia hypotension, Obstructive shock, and LV shock. Table 9 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 FER produced the labels (ground truth) for training the ML algorithm.

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

Table 9 - The full expert-based rules that make the research's ground truth.

### 3.6. Expert-Based Rules Evaluation

For initial model evaluation, a simple analysis was performed on a dataset that contained complete samples; where a "complete sample" means: the whole data for all required parameters (HR, ARTBPS, ARTBPM, CVP, PAPD, RR Total and RR mandatory) per patient's diagnosis at a certain minute. The dataset

contained 6,917 recording minutes and consisted of 4,904 monitor alarms in total.

FER classification results on the complete samples were compared to the classification that is made by the current TAMC monitor. The average of a monitor's alarms ranges from 82-115 per day while 3-5 alarms occurred per hour per subject. Of these, 43.4% were categorized as FA by the proposed FER. The following tables (Table 10-11) show the monitor and FER alarms' percentages:

| Type           | Amount | %   |
|----------------|--------|-----|
| Monitor Alarms | 4,904  | 70  |
| No Alarm       | 2,013  | 30  |
| Total          | 6,917  | 100 |

Table 10 - Percentages of monitor alarms.

| Type       | Amount | %     |
|------------|--------|-------|
| FER Alarms | 1,962  | 28.36 |
| No Alarm   | 4,955  | 71.63 |
| Total      | 6,917  | 100   |

Table 11 - Percentages of FER alarms.

A comparison between the monitor alarms and FER is presented in the confusion matrix below (Table 12) while FER' classification was used as the ground truth.

| FER \ Monitor | Positive    | Negative    |
|---------------|-------------|-------------|
| Positive      | TP – 27.49% | FN – 0.86%  |
| Negative      | FP – 43.4%  | TN – 28.23% |

Table 12 - Confusion matrix of monitor performances.

The algorithm employed by the monitors was shown to have a sensitivity of 96.94% with a specificity of 39.41% and precision of 38.78%. Additionally, classification of complete samples data by FER demonstrated the potential for using multiple physiologic parameters to reduce the FAR in the clinical setting.

Second analysis compared FER classification on the complete samples and PER over the partial parameters set (in cases of up to 3 missing parameters), i.e. without taking into consideration logical clauses relying on missing parameters in the current partial dataset. For example, the FER contains a rule stating that if ARTBPM is smaller than 50 mm Hg, and CVP is either smaller than 5 or larger than -10 mm Hg, an alarm should be triggered. For the alternative dataset in which the ARTBPM is missing, the rule was reduced to an alarm being triggered whether CVP is either smaller than 5 or larger than -10 mm Hg, without the clause regarding the ARTBPM.

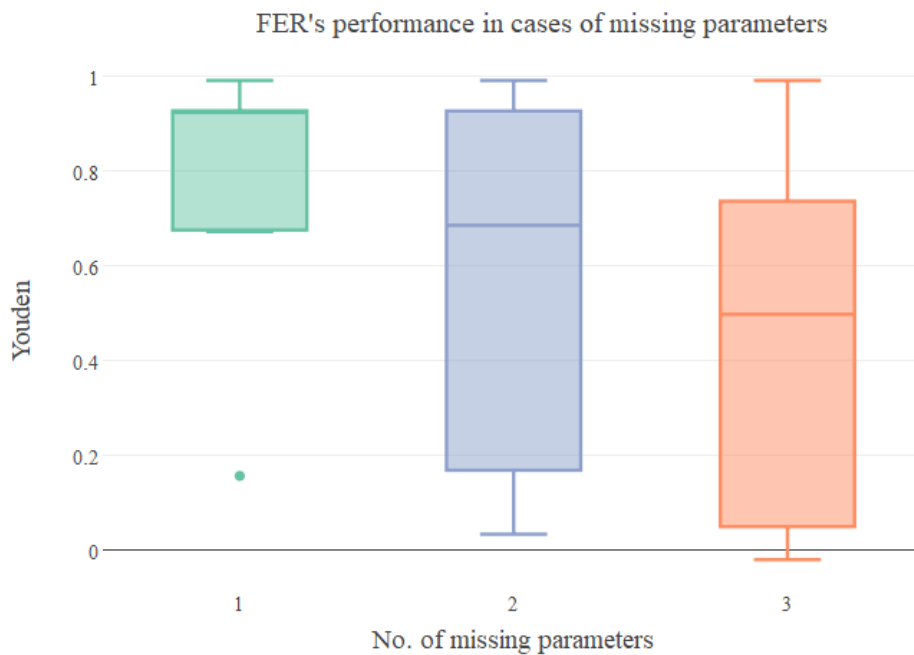


Figure 9 – FER' performance in cases of missing parameters.

As presented in the graph above (Figure 9), the classification of PER (with absence of parameters) results in poor performances which grows as function of missing parameters and thus leads to less efficiency in diagnosing patient's condition (see [Appendix D](#) for full analysis).

### 3.7. Methodology implementation

As presented above, FER had some disabilities in certain situations that need a special treatment. FER employ physiological parameters that are sampled by



different sensors; however, in 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, a dataset that contains 20,045 samples without missing values was extracted from the clinical database. Each such sample portrays simultaneous measuring of twelve physiological parameters for a single patient at a specific minute. The extracted parameters are: HR, ARTBPS, ARTBPM, CVP, PAPD, RR total, RR mandatory, Spo2, Fio2, ST1, ST2, ST3. Some of the parameters were chosen because of their medical importance (as part of the FER presented in Table 9) and some because of their high prevalence in the ICU recordings.

In order to investigate situations of all the clinical alarm scenarios combined and each one individually, two tests were conducted. A dataset that contains at least 1,500 events of each of the alarm scenarios was extracted for the first test. This number of events was chosen according to the maximum number of events in the smallest dataset, out of the seven extracted clinical scenarios datasets. In the second test, samples of the seven alarm scenarios were divided into seven separate datasets. The description of the following steps is relevant for both tests.

To minimize bias and variance in the learning phase, the datasets were divided according to the ML methodology of cross validation (Bishop, 2007) into three approximately equal sized subsets, which alternatively used for training, validation, 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 the Youden's index. The output of a model provided the estimated probability for alarm triggering for each of the twelve-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 it (or equal) 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 results (estimating the accuracy in alarm classification in the "future"), an average was calculated over the test results of the three test subsets. These averages are presented in the Results section. Using the Youden's index while training and validating the candidate RF models, and not only in evaluating (testing) the models, guarantees that the selected RF, having specific numbers of trees and parameters, is augmented from the outset toward maximizing both sensitivity and specificity.

As mentioned before, FER model 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 (FER with no missing parameters). This allowed us to compare the decisions made by a classifier set by a medical expert (FER\PER) and an automated ML model that is data-driven, mimicking the expert decision making.

The tests were implemented using the R package randomForest (Liaw & Wiener, 2002) using Rstudio software version 3.4.1. The source code is available on the GitHub repository: <https://github.com/galhev/Thesis>.

## 4. Results

Table 13 presents the average performance measures in the first test (all the clinical alarm scenarios together) for models with up to three missing parameters.

| # of Missing Parameters | Model | Youden's Index | Precision | Specificity | Sensitivity | %FP |
|-------------------------|-------|----------------|-----------|-------------|-------------|-----|
| None                    | RF    | 0.99           | 0.99      | 0.99        | 0.99        | 1%  |
|                         | FER   | 1              | 1         | 1           | 1           | 0%  |
| One                     | RF    | 0.97           | 0.99      | 0.99        | 0.98        | 1%  |
|                         | PER   | 0.8            | 0.88      | 0.83        | 0.97        | 17% |
| Two                     | RF    | 0.96           | 0.98      | 0.98        | 0.97        | 2%  |
|                         | PER   | 0.66           | 0.81      | 0.71        | 0.95        | 29% |
| Three                   | RF    | 0.94           | 0.98      | 0.98        | 0.95        | 2%  |
|                         | PER   | 0.54           | 0.76      | 0.61        | 0.92        | 39% |

*Table 13 - Average models performance with and without missing parameters. Note that PER for no missing data is the FER, which is the reference/benchmark result of this research.*

Table 13 shows that in the first test, while none of the parameters was missing, the difference between the RF model scores and FER was negligible. Importantly, the absolute rate of FA was low for one (1%), two (1%) and three (2%) missing parameters compared to PER (17%, 29%, 39%). Additionally, the expanded results of Table 13 (see Appendix E) shows a clear superiority of RF compare to PER. In all the cases of missing parameters (one, two and three), RF achieved significantly better score, equal or less than PER with negligible difference (1%) in all the measures (youden's index, precision, specificity, sensitivity and FP). Out of 63 combinations of tests in total of one (7), two (21), and three (35) missing parameters, only in 4 (6.34%) cases PER achieved better score than RF, 6 cases (9.5%) in which RF and PER achieved equal scores and 53 cases (84%) in which RF was superior compare to PER.

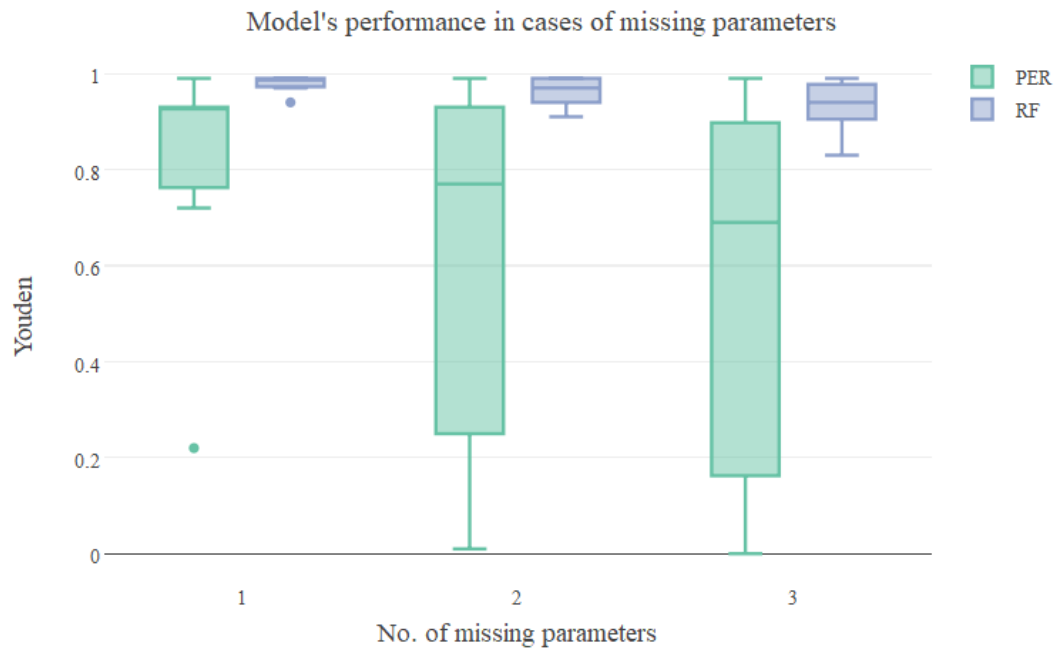


Figure 10 - Models' Youden's indexes for missing parameters.

Figure 10 (also available in <http://rpubs.com/galhev/377607>) presents the expanded results in Appendix A in a boxplot. The graph shows a clear trend for PER performance, which gets dramatically poorer as more parameters are missing from the data, in contrast with the RF that remains high sensitivity and specificity.

The second test included seven different data sets, each contained samples for a different alarm scenario. Table 14 presents average performance measures for each model in each clinical alarm scenario for a different number of missing parameters used to compose the rules, each time the missing parameters were those yielding the poorest model performance.

| Scenario          | # of alarm events  | Missing Parameters  | Model | Youden's Index | Precision | Specificity | Sensitivity | FP  |
|-------------------|--------------------|---------------------|-------|----------------|-----------|-------------|-------------|-----|
| 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         | 17% |
|                   |                    |                     | PER   | 0.25           | 0.59      | 0.25        | 1           | 75% |
|                   |                    | CVP                 | RF    | 0.99           | 0.99      | 0.99        | 0.99        | 1%  |
|                   |                    |                     | PER   | 0.97           | 0.99      | 0.98        | 0.99        | 2%  |
|                   |                    | PAPD                | RF    | 0.99           | 0.99      | 0.98        | 0.99        | 2%  |
|                   |                    |                     | PER   | 0.86           | 0.93      | 0.89        | 0.97        | 11% |
|                   |                    | ARTBPS + CVP        | RF    | 0.76           | 0.65      | 0.46        | 0.96        | 54% |
|                   |                    |                     | PER   | 0.21           | 0.52      | 0.2         | 0.95        | 80% |
|                   |                    | ARTBPS + PAPD       | RF    | 0.91           | 0.86      | 0.84        | 0.95        | 16% |
|                   |                    |                     | PER   | 0.04           | 0.54      | 0.04        | 1           | 96% |
|                   |                    | CVP + PAPD          | RF    | 0.98           | 0.87      | 0.78        | 0.9         | 22% |
|                   |                    |                     | PER   | 0.86           | 0.9       | 0.87        | 0.97        | 13% |
|                   |                    | ARTBPS + CVP + PAPD | RF    | 0.64           | 0.71      | 0.61        | 0.64        | 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        | 17% |
|                   |                    |                     | PER   | 0.25           | 0.8       | 0.28        | 0.97        | 72% |
|                   |                    | CVP                 | RF    | 0.99           | 0.99      | 0.97        | 1           | 3%  |
|                   |                    |                     | PER   | 0.98           | 0.99      | 0.98        | 1           | 2%  |
|                   |                    | PAPD                | RF    | 0.97           | 0.98      | 0.95        | 1           | 5%  |
|                   |                    |                     | PER   | 0.85           | 0.89      | 0.9         | 0.95        | 10% |
|                   |                    | ARTBPS + CVP        | RF    | 0.71           | 0.82      | 0.77        | 0.82        | 23% |
|                   |                    |                     | PER   | 30.2           | 0.68      | 0.48        | 0.75        | 52% |
|                   |                    | ARTBPS + PAPD       | RF    | 0.86           | 0.91      | 0.91        | 0.87        | 9%  |
|                   |                    |                     | PER   | 0.02           | 0.57      | 0.02        | 1           | 98% |
|                   |                    | CVP + PAPD          | RF    | 1              | 1         | 1           | 1           | 0%  |
|                   |                    |                     | PER   | 0.59           | 0.78      | 0.77        | 0.82        | 23% |
|                   |                    | ARTBPS + CVP + PAPD | RF    | 0.68           | 0.85      | 0.55        | 0.9         | 45% |
|                   |                    |                     | PER   | -              | -         | -           | -           | -   |

| Scenario                | # of alarm events  | Missing Parameters | Model | Youden's Index | Precision | Specificity | Sensitivity | FP  |
|-------------------------|--------------------|--------------------|-------|----------------|-----------|-------------|-------------|-----|
| 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        | 2%  |
|                         |                    |                    | PER   | 0.7            | 0.92      | 0.88        | 0.82        | 12% |
|                         |                    | ARTBPS             | RF    | 0.98           | 1         | 1           | 0.98        | 0%  |
|                         |                    |                    | PER   | 0.39           | 0.58      | 0.39        | 1           | 61% |
|                         |                    | HR + ARTBPS        | RF    | 0.96           | 0.87      | 0.78        | 0.99        | 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        | 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           | 1%  |
|                         |                    |                    | PER   | 0.78           | 0.85      | 0.78        | 1           | 22% |
|                         |                    | ARTBPS             | RF    | 0.99           | 1         | 1           | 0.98        | 0%  |
|                         |                    |                    | PER   | 0.27           | 0.62      | 0.27        | 1           | 73% |
|                         |                    | HR + ARTBPS        | RF    | 0.78           | 0.72      | 0.65        | 0.95        | 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        | 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        | 3%  |
|                         |                    |                    | 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        | 45% |
|                         |                    |                    | PER   | -              | -         | -           | -           | -   |

Table 14 - Averages model performance measures for different clinical alarm scenario and missing parameter setting. Note: Green/red cells indicate higher/lower scores in comparing RF and PER for each of the missing parameters in each clinical alarm scenario. An empty cell indicates the inability to exercise the PER model due to missing parameter(s) this model is based on.

Unlike the RF model, the FER/PER classification results are based on a single value (binary output obtained by the model rules, e.g., a sample with values of ARTBPM < 50 mm Hg , CVP < 5 mm Hg, CVP > -10 mm Hg will indicate about Hypovolemia scenario and the output will be 1); thus, the calculation of Youden's index is also based on a single value (without maximizing c value). Missing results of PER (indicated by “-“) in Table 14 reflect situations of inability to exercise the 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, e.g., if the HR parameter is missing, it is impossible to exercise the single rule, HR<45, which is needed to detect Bradycardia. In such situations, the RF model – integrating and fusing information from the available parameters compensating for those missing – has a significant advantage over the PER model that is not applicable. As presented for example in Table 14, looking at the Hypovolemia scenario, when all the parameters which composed the rules were missing (ARTBPM, CVP), RF still succeeded to produce good results with Youden's index and precision measure of 0.8 and 0.7, respectively, using the remaining parameters, while PER failed.

According to Table 14, the Obstructive shock FAR were reduced between 1% to 80% by RF when one and two parameters were missing, except the case where CVP + PAPD were missing and PER achieved better score (13%) than RF (22%). When all three parameters used to compose the rules of Obstructive shock were missing, PER failed while RF achieved FAR of 39%. The scores of PER and RF were equal to Bradycardia and Tachycardia except in the case where HR was missing. In that case, PER failed and RF achieved FAR of 41% and 59% respectively. Bradycardia hypotension and Tachycardia hypotension FAR were reduced between 10% to 61% and 21% to 73% respectively by RF when one parameter was missing. When all two parameters used to compose the rules of Bradycardia hypotension and Tachycardia hypotension were missing, FAR were 22% and 35% respectively by RF while PER failed. The false LV shock alarm suppression rate by RF was the highest of all alarm scenarios tested, with a reduction in FAR down to 89% except in the case where CVP was missing. In that case, PER achieved better score (2%) than RF (3%); however, the difference between the results was negligible. In the case where all the parameters used to compose the rules were missing, PER failed while RF achieved FAR of 45%. The Hypovolemia

FAR were equal to RF and PER except in the case where ARTBPM was missing. In that case, PER achieved better score (0%) than RF (3%). When all the parameters used to compose the rules of Hypovolemia were missing PER failed while RF achieved FAR of 45%. In both the first (Table 13) and second (Table 14) tests, RF presents a significant advantage over PER model and a negligible difference with FER (the ground truth).

The most important parameters for classifying each scenario were examined. Table 15 presents the parameters that achieved the highest ranking according to the Gini index (Breiman et al., 1984). For example, when none of the parameters used to compose the rules of Hypovolemia is missing, ARTBPM achieved the highest reduction in impurity for classification the dataset. Then, when ARTBPM was missing (column "One"), the parameter that achieved the highest reduction in impurity was CVP and so on and so forth.

| <b># of Missing Parameters</b><br><b>Clinical Alarm scenario</b> | <b>No</b> | <b>One</b> | <b>Two</b> | <b>Three</b> |
|--|-----------|------------|------------|--------------|
| <b>Hypovolemia</b>   | ARTBPM    | CVP        | ARTBPS     | PAPD         |
| <b>LV shock</b>  | ARTBPS    | ARTBPM     | PAPD       | CVP          |
| <b>Obstructive shock</b>   | ARTBPS    | CVP        | ARTBPM     | PAPD         |
| <b>Bradycardia hypotension</b>                                   | ARTBPS    | HR         | ARTBPM     | ST1          |
| <b>Tachycardia hypotension</b>                                   | HR        | ARTBPS     | ARTBPM     | ST1          |
| <b>Bradycardia</b>   | HR        | ARTBPS     | ST1        | RR mandatory |
| <b>Tachycardia</b>   | HR        | ARTBPS     | ST1        | RR mandatory |

Table 15 - Parameter importance ranking according to the Gini index.

Not surprisingly, the highest scores were achieved when the parameters used to compose the rules were not missing in the dataset. However, the increase in the number of missing source parameters, resulted in the model forecast constructed from the remaining parameters, compensating for the missingness. According to Table 15, the most important parameters to classify most of the alarms scenarios (when none of the parameters were missing) were HR and ARTBPS. Additionally, Table 1 shows that RF detected correctly the parameters used to compose the rules of all the clinical alarm



scenarios. Interesting parameters are discovered when all the parameters used to compose the rules of a particular scenario were missing. For example, in the case when HR was missing in Tachycardia, the next parameter that compensate on the missingness was ARTBPS. This compensation make sense in light of the fact that both related to the contractions of the heart. When ARTBPS was also missing in Tachycardia, RF managed to detect the potential of ST1 (which also presents contractions of the heart) to compensate on the missingness of the previous two (HR and ARTBPS).

## 5. Discussion

Medical monitors are a vital part of any health care facility. They allow medical professionals to assess and keep track of a patient's condition and progress while under their care over time. Medical monitors in use today in the ICU diagnose a patient's status by analyzing each vital sign separately. Unfortunately, the high FAR of clinical monitors is staggering.

Previous investigations towards FAR reduction have presented an approach that is driven by expert decision assessing the patient's state using multiple sensors instead of just one of them. This approach proved to produce more sophisticated rules of alarm scenarios than do present single-sensor based systems and thus have enormous potential to reduce the FAR. However, this approach presents poor performance in situations of missing sensor data, which is common in the ICU. The greater the missingness of sensor data, the less accurate the approach becomes, presenting similar performance to that of single-sensor based monitoring systems (assuming that sensor is not missing).

This study have improved this approach for cases of missing data by applying an ML method to multi-sensor information. RF algorithm was used 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. The results demonstrate that based on several accuracy performance measures the efficiency of the RF model in overcoming missing sensor data is appreciable. Hence, the suggested approach is a simple and practical solution for suppressing FAR significantly in situations of missing sensor data, and reducing alarm fatigue in ICUs.

## 6. Conclusions

The technology for building knowledge-based systems by inductive inference from examples has been demonstrated successfully in several practical applications. This work presented an approach for FAR reduction, which 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.

This research showed that the data-driven learning approach manages to recreate expert alarm classification, even when some sensor (parameter) data are missing. The system is accurate in imitating the expert classification, thus, its performance is comparable to that of the human medical expert. In addition, in cases of data loss of one or more physiological parameters, relying on the ML-based approach to exploit information from the partial parameter set, enables it to handle the classification task significantly better than the expert rules relying on the partial data. Thus, it might be considered as an alternative to the existing clinical alarm system.

In conclusion, the approach presented in this research 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 an informed decision. As demonstrated, the 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 scarifying the FAR. Note also that this approach allows an operating point to be selected for each individual patient by changing the thresholds set for the expert rules to balance the desired sensitivity and specificity for this 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 aiding clinicians.

## **7. Merit of the research and proposed contribution to science**

Attempts to reduce FAR was made in previous investigations but none of these approaches advanced into the mainstream of patient monitoring. This research tried to present an easy and practical way to improve the current medical monitor by clinical alarm scenarios that could be programmed into the bedside monitors.

This research introduce new rules which simulate an experienced expert doctor who diagnoses a patient's condition and propose a solution for handling with lack or faulty sensors causing poor diagnosing performances. The proposed approach can potentially assist caregivers in predicting clinical alarm scenarios, reducing alarm burden, providing a complementary tool to support clinical decision-making, and enhancing patient monitoring. Additionally, the presented FER enable processing of clinical information in a similar manner to human operators. This approach has the potential to significantly reduce FA, increase the positive predictive value of alarms, and make some progress reducing the ubiquitous problem of alarm fatigue in the ICU.

## **8. Limitations and future research**

This study encompasses several limitations that should be mentioned. In the critical care setting, it is difficult to record everything that can be potentially useful in retrospective research. Thus, parameter values were collected based on assumptions regarding future research needs. In retrospect, additional clinical information was needed as well as different sampling frequencies to reconstruct the clinical context necessary in interpreting a past scenario. This study promotes such and similar future efforts by providing guidelines and a methodology for ICU monitoring that wishes to enjoy the benefits of the approach presented in this research.

Key issue of the present study is that tagging of samples was based solely on the FER and was not validated by a doctor because of lack of resources. It is unlikely that a large number of significant events were missed, but precise estimation of the performance of these rules would require this more reliable database. Additionally, the FER were written only by one specialist doctor and were not reviewed by other doctors, a topic that should be addressed in future studies.

A further investigation that should be carried out in follow-up studies is an alarm type prediction with missing parameters based on expert knowledge instead of a binary classification.

## 9. Appendixes

### 9.1. Appendix A – Classification Terms

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}$$

$$\text{Positive predictive value} = \text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

$$\text{Negative predictive value} = \frac{\text{True negatives}}{\text{True negatives} + \text{False negatives}}$$

Altman & Bland (1994).

### 9.2. Appendix B – Descriptive Statistics

|         | HR    | ARTBPS | ARTBPM | CVP   | PAPD  | RR total | RR mandatory | Spo2  | ST1   | ST2   | ST3   | Fio2  |
|---------|-------|--------|--------|-------|-------|----------|--------------|-------|-------|-------|-------|-------|
| Min.    | 0     | -9     | -15    | -39   | -34   | 3        | 0            | 44    | -327  | -327  | -327  | 30    |
| 1st Qu. | 77    | 87     | 58     | 8     | 15    | 14       | 12           | 95    | -0.36 | -0.25 | -0.42 | 40    |
| Median  | 93    | 106    | 69     | 10    | 19    | 16       | 14           | 97    | 0     | 0.07  | -0.04 | 50    |
| Mean    | 94.68 | 110    | 72.88  | 11.93 | 19.91 | 15.9     | 12.88        | 95.73 | -3.4  | -3.3  | -3.51 | 52.28 |
| 3rd Qu. | 113   | 129    | 83     | 14    | 24    | 18       | 16           | 99    | 0.37  | 0.62  | 0.2   | 60    |
| Max.    | 238   | 309    | 313    | 314   | 301   | 128      | 26           | 100   | 9     | 55.47 | 30.14 | 100   |

The main goal of the data visualization tool, which was developed in this research, was a convenient way for data reviewing and classification by an expert annotator. The reviewers would be able to view all the signals surrounding each alarm (with a controllable window size). They could also expand and shrink the time window at their discretion to provide more detailed information or to mark each alarm as true or false.

One of the most important limitations of the present study is the inability to find an expert who would review and classify the dataset, thus the presented tool was not used apart from building the descriptive statistics. Hopefully that future studies will use the tool for this purpose.

### 9.3. Appendix C – Alarm Frequency

#### Morning Shifts:

| Weekday | Mean   | Median | SD    | Max | Min |
|---------|--------|--------|-------|-----|-----|
| 1       | 248.33 | 241    | 88.09 | 571 | 3   |
| 2       | 250.94 | 246    | 87.54 | 589 | 13  |
| 3       | 248.86 | 242    | 84.33 | 594 | 10  |
| 4       | 250.87 | 245    | 87.37 | 636 | 3   |
| 5       | 251.37 | 241    | 85.70 | 695 | 1   |
| 6       | 252.22 | 247    | 86.16 | 605 | 28  |
| 7       | 254.33 | 247    | 87.46 | 600 | 1   |

#### Afternoon Shifts:

| Weekday | Mean   | Median | SD    | Max | Min |
|---------|--------|--------|-------|-----|-----|
| 1       | 244.46 | 237    | 88.75 | 587 | 20  |
| 2       | 240.31 | 229.5  | 87.47 | 635 | 14  |
| 3       | 243.90 | 236    | 84.12 | 534 | 1   |
| 4       | 245.58 | 239    | 85.12 | 567 | 2   |
| 5       | 248.83 | 243    | 83.39 | 531 | 3   |
| 6       | 247.02 | 239    | 89.17 | 584 | 4   |
| 7       | 247.67 | 240    | 85.86 | 624 | 1   |

#### Night Shifts:

| Weekday | Mean   | Median | SD    | Max | Min |
|---------|--------|--------|-------|-----|-----|
| 1       | 259.89 | 252    | 88.81 | 601 | 1   |
| 2       | 260.92 | 253    | 90.58 | 625 | 2   |
| 3       | 260.37 | 257    | 87.91 | 656 | 2   |
| 4       | 263.69 | 256    | 87.97 | 586 | 2   |
| 5       | 265.06 | 256    | 89.33 | 555 | 3   |
| 6       | 264.49 | 257    | 89.32 | 614 | 13  |
| 7       | 260.82 | 253    | 89.24 | 632 | 1   |

## 9.4. Appendix D – FER performances in cases of missing parameters

### One parameter is missing:

| Missing Parameter | Sensitivity | Specificity | Precision | Youden's Index |
|-------------------|-------------|-------------|-----------|----------------|
| HR                | 0.89        | 0.79        | 0.87      | 0.68           |
| ARTBPS            | 0.99        | 0.16        | 0.66      | 0.15           |
| ARTBPM            | 0.94        | 0.73        | 0.68      | 0.67           |
| CVP               | 1           | 0.99        | 0.98      | 0.99           |
| PAPD              | 0.94        | 0.98        | 0.97      | 0.92           |
| RR total          | 0.94        | 0.98        | 0.97      | 0.93           |
| RR mandatory      | 0.94        | 0.98        | 0.97      | 0.93           |

### Two parameters are missing:

| Missing Parameters      | Sensitivity | Specificity | Precision | Youden's Index |
|-------------------------|-------------|-------------|-----------|----------------|
| HR + ARTBPS             | 0.12        | 0.96        | 0.64      | 0.08           |
| HR + ARTBPM             | 0.79        | 0.7         | 0.62      | 0.5            |
| HR + CVP                | 0.82        | 0.89        | 0.81      | 0.71           |
| HR + PAPD               | 0.79        | 0.89        | 0.81      | 0.68           |
| HR + RR total           | 0.79        | 0.89        | 0.81      | 0.68           |
| HR + RR mandatory       | 0.79        | 0.89        | 0.81      | 0.68           |
| ARTBPS + ARTBPM         | 0.87        | 0.16        | 0.63      | 0.03           |
| ARTBPS + CVP            | 1           | 0.17        | 0.67      | 0.17           |
| ARTBPS + PAPD           | 0.99        | 0.04        | 0.63      | 0.04           |
| ARTBPS + RR total       | 0.99        | 0.16        | 0.66      | 0.16           |
| ARTBPS + RR mandatory   | 0.99        | 0.16        | 0.66      | 0.16           |
| ARTBPM + CVP            | 0.78        | 1           | 1         | 0.78           |
| ARTBPM + PAPD           | 0.79        | 0.94        | 0.96      | 0.74           |
| ARTBPM + RR total       | 0.73        | 0.94        | 0.95      | 0.67           |
| ARTBPM +RR mandatory    | 0.73        | 0.94        | 0.95      | 0.67           |
| CVP + PAPD              | 0.99        | 1           | 1         | 0.99           |
| CVP + RR total          | 0.99        | 1           | 1         | 0.99           |
| CVP + RR mandatory      | 0.99        | 1           | 1         | 0.99           |
| PAPD + RR total         | 0.98        | 0.94        | 0.97      | 0.93           |
| PAPD + RR mandatory     | 0.98        | 0.94        | 0.97      | 0.93           |
| RR total + RR mandatory | 0.98        | 0.94        | 0.97      | 0.93           |



**Three parameters are missing:**

| Missing Parameters                      | Sensitivity | Specificity | Precision | Youden's Index |
|---|-------------|-------------|-----------|----------------|
| <b>HR + ARTBPS + ARTBPM</b>             | 0.86        | 0.12        | 0.62      | -0.02          |
| <b>HR + ARTBPS + CVP</b>                | 0.96        | 0.13        | 0.65      | 0.09           |
| <b>HR + ARTBPS + PAPD</b>               | 0.98        | 0.03        | 0.63      | 0.01           |
| <b>HR + ARTBPS + RR total</b>           | 0.89        | 0.16        | 0.64      | 0.05           |
| <b>HR + ARTBPS + RR mandatory</b>       | 0.89        | 0.16        | 0.64      | 0.05           |
| <b>HR + ARTBPM + CVP</b>                | 0.82        | 0.67        | 0.6       | 0.49           |
| <b>HR + ARTBPM + PAPD</b>               | 0.79        | 0.7         | 0.62      | 0.5            |
| <b>HR + ARTBPM + RR total</b>           | 0.79        | 0.7         | 0.62      | 0.5            |
| <b>HR + ARTBPM + RR mandatory</b>       | 0.79        | 0.7         | 0.62      | 0.5            |
| <b>HR + CVP + PAPD</b>                  | 0.82        | 0.89        | 0.81      | 0.71           |
| <b>HR + CVP + RR total</b>              | 0.82        | 0.89        | 0.81      | 0.71           |
| <b>HR + CVP + RR mandatory</b>          | 0.82        | 0.89        | 0.81      | 0.71           |
| <b>HR + PAPD + RR total</b>             | 0.79        | 0.89        | 0.81      | 0.68           |
| <b>HR + PAPD + RR mandatory</b>         | 0.79        | 0.89        | 0.81      | 0.68           |
| <b>HR + RR total + RR mandatory</b>     | 0.79        | 0.89        | 0.81      | 0.68           |
| <b>ARTBPS + ARTBPM + CVP</b>            | 0.83        | 0.17        | 0.63      | 0.01           |
| <b>ARTBPS + ARTBPM + PAPD</b>           | 0.97        | 0.04        | 0.63      | 0.01           |
| <b>ARTBPS + ARTBPM + RR total</b>       | 0.87        | 0.16        | 0.63      | 0.03           |
| <b>ARTBPS + ARTBPM + RR mandatory</b>   | 0.87        | 0.16        | 0.63      | 0.03           |
| <b>ARTBPS + CVP + PAPD</b>              | 0.9         | 0.89        | 0.83      | 0.79           |
| <b>ARTBPS + CVP + RR total</b>          | 1           | 0.17        | 0.67      | 0.17           |
| <b>ARTBPS + CVP + RR mandatory</b>      | 0.97        | 0.18        | 0.66      | 0.15           |
| <b>ARTBPS + PAPD + RR total</b>         | 0.99        | 0.04        | 0.63      | 0.04           |
| <b>ARTBPS + PAPD + RR mandatory</b>     | 0.99        | 0.04        | 0.63      | 0.04           |
| <b>ARTBPS + RR total + RR mandatory</b> | 0.99        | 0.16        | 0.66      | 0.16           |
| <b>ARTBPM + CVP + PAPD</b>              | 1           | 0.78        | 0.73      | 0.78           |
| <b>ARTBPM + CVP + RR total</b>          | 1           | 0.78        | 0.73      | 0.78           |
| <b>ARTBPM + CVP + RR mandatory</b>      | 1           | 0.78        | 0.73      | 0.78           |
| <b>ARTBPM + PAPD + RR total</b>         | 0.94        | 0.79        | 0.73      | 0.74           |
| <b>ARTBPM + PAPD + RR mandatory</b>     | 0.94        | 0.79        | 0.73      | 0.74           |
| <b>ARTBPM + RR total + RR mandatory</b> | 0.94        | 0.73        | 0.68      | 0.67           |
| <b>CVP + PAPD + RR total</b>            | 1           | 0.99        | 0.98      | 0.99           |
| <b>CVP + PAPD + RR mandatory</b>        | 1           | 0.99        | 0.98      | 0.99           |
| <b>CVP + RR total + RR mandatory</b>    | 1           | 0.99        | 0.98      | 0.99           |
| <b>PAPD + RR total + RR mandatory</b>   | 0.94        | 0.98        | 0.97      | 0.93           |

## 9.5. Appendix E – Full Results

### One parameter is missing:

| Missing Parameter   | Model      | Youden's Index | Precision   | Specificity | Sensitivity | %FP        | AUC  |
|---------------------|------------|----------------|-------------|-------------|-------------|------------|------|
| <b>HR</b>           | RF         | 0.94           | 0.99        | 0.99        | 0.95        | 0.01       | 0.99 |
|                     | PER        | 0.72           | 0.84        | 0.84        | 0.89        | 16%        |      |
| <b>ARTBPS</b>       | RF         | 0.97           | 0.99        | 0.99        | 0.98        | 0.01       | 1    |
|                     | PER        | 0.22           | 0.56        | 0.22        | 1           | 78%        |      |
| <b>ARTBPM</b>       | RF         | 0.98           | 0.99        | 0.99        | 0.99        | 0.01       | 1    |
|                     | PER        | 0.89           | 0.94        | 0.94        | 0.95        | 6%         |      |
| <b>CVP</b>          | RF         | 0.99           | 1           | 1           | 0.99        | 0          | 1    |
|                     | PER        | 0.99           | 1           | 1           | 0.99        | 0          |      |
| <b>PAPD</b>         | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1    |
|                     | PER        | 0.93           | 0.94        | 0.94        | 0.99        | 6%         |      |
| <b>RR total</b>     | RF         | 0.99           | 1           | 1           | 0.99        | 0          | 1    |
|                     | PER        | 0.93           | 0.94        | 0.94        | 0.99        | 6%         |      |
| <b>RR mandatory</b> | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1    |
|                     | PER        | 0.93           | 0.94        | 0.94        | 0.99        | 6%         |      |
| <b>Average</b>      | <b>RF</b>  | <b>0.98</b>    | <b>1.00</b> | <b>1.00</b> | <b>0.98</b> | <b>0%</b>  |      |
|                     | <b>PER</b> | <b>0.80</b>    | <b>0.88</b> | <b>0.83</b> | <b>0.97</b> | <b>17%</b> |      |

*Note: Unlike the RF model, the FER/PER classification results are based on a single value (binary output obtained by the model rules, e.g., a sample with values of ARTBPM < 50 mm Hg , CVP < 5 mm Hg, CVP > -10 mm Hg will indicate about Hypovolemia scenario and the output will be 1);thus, the calculation of Youden's index is also based on a single value. That is, it is impossible to calculate the value of AUC for PER; therefore these cells are empty in the table for this model.*

**Two parameters are missing:**

| Missing Parameters           | Model | Youden's Index | Precision | Specificity | Sensitivity | %FP | AUC  |
|------------------------------|-------|----------------|-----------|-------------|-------------|-----|------|
| <b>HR + ARTBPS</b>           | RF    | 0.91           | 0.97      | 0.97        | 0.93        | 3%  | 0.99 |
|                              | PER   | 0.06           | 0.51      | 0.1         | 0.95        | 90% |      |
| <b>HR + ARTBPM</b>           | RF    | 0.93           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                              | PER   | 0.69           | 0.84      | 0.84        | 0.86        | 16% |      |
| <b>HR + CVP</b>              | RF    | 0.94           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                              | PER   | 0.77           | 0.89      | 0.89        | 0.87        | 11% |      |
| <b>HR + PAPD</b>             | RF    | 0.94           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                              | PER   | 0.72           | 0.84      | 0.84        | 0.89        | 16% |      |
| <b>HR + RR total</b>         | RF    | 0.94           | 0.98      | 0.99        | 0.95        | 1%  | 0.99 |
|                              | PER   | 0.72           | 0.84      | 0.84        | 0.89        | 16% |      |
| <b>HR + RR mandatory</b>     | RF    | 0.94           | 0.98      | 0.99        | 0.95        | 1%  | 0.99 |
|                              | PER   | 0.72           | 0.84      | 0.84        | 0.89        | 16% |      |
| <b>ARTBPS + ARTBPM</b>       | RF    | 0.91           | 0.97      | 0.97        | 0.94        | 3%  | 0.99 |
|                              | PER   | 0.2            | 0.55      | 0.22        | 0.97        | 78% |      |
| <b>ARTBPS + CVP</b>          | RF    | 0.97           | 0.99      | 0.99        | 0.98        | 1%  | 1    |
|                              | PER   | 0.26           | 0.57      | 0.27        | 1           | 73% |      |
| <b>ARTBPS + PAPD</b>         | RF    | 0.97           | 0.99      | 0.99        | 0.98        | 1%  | 1    |
|                              | PER   | 0.01           | 0.5       | 0.01        | 1           | 99% |      |
| <b>ARTBPS + RR total</b>     | RF    | 0.97           | 0.99      | 0.99        | 0.98        | 1%  | 1    |
|                              | PER   | 0.22           | 0.56      | 0.22        | 1           | 78% |      |
| <b>ARTBPS + RR mandatory</b> | RF    | 0.97           | 0.99      | 0.99        | 0.98        | 1%  | 1    |
|                              | PER   | 0.22           | 0.56      | 0.22        | 1           | 78% |      |
| <b>ARTBPM + CVP</b>          | RF    | 0.98           | 0.99      | 0.99        | 0.99        | 1%  | 1    |
|                              | PER   | 0.9            | 1         | 1           | 0.9         | 0%  |      |
| <b>ARTBPM + PAPD</b>         | RF    | 0.98           | 0.99      | 0.99        | 0.99        | 1%  | 1    |
|                              | PER   | 0.9            | 0.94      | 0.94        | 0.96        | 6%  |      |
| <b>ARTBPM + RR total</b>     | RF    | 0.98           | 0.99      | 0.99        | 0.99        | 1%  | 1    |
|                              | PER   | 0.89           | 0.94      | 0.94        | 0.95        | 6%  |      |
| <b>ARTBPM + RR mandatory</b> | RF    | 0.98           | 0.99      | 0.99        | 0.99        | 1%  | 1    |
|                              | PER   | 0.89           | 0.94      | 0.94        | 0.95        | 6%  |      |
| <b>CVP + PAPD</b>            | RF    | 0.99           | 1         | 1           | 0.99        | 0%  | 1    |
|                              | PER   | 0.99           | 1         | 1           | 0.99        | 0%  |      |
| <b>CVP + RR total</b>        | RF    | 0.99           | 1         | 1           | 0.99        | 0%  | 1    |
|                              | PER   | 0.99           | 1         | 1           | 0.99        | 0%  |      |
| <b>CVP + RR mandatory</b>    | RF    | 0.99           | 1         | 1           | 0.99        | 0%  | 1    |
|                              | PER   | 0.99           | 1         | 1           | 0.99        | 0%  |      |
| <b>PAPD + RR total</b>       | RF    | 0.99           | 1         | 1           | 0.99        | 0%  | 1    |
|                              | PER   | 0.93           | 0.94      | 0.94        | 0.99        | 6%  |      |

| Missing Parameters             | Model      | Youden's Index | Precision   | Specificity | Sensitivity | %FP        | AUC |
|--------------------------------|------------|----------------|-------------|-------------|-------------|------------|-----|
| <b>PAPD + RR mandatory</b>     | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1   |
|                                | PER        | 0.93           | 0.94        | 0.94        | 0.99        | 6%         |     |
| <b>RR total + RR mandatory</b> | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1   |
|                                | PER        | 0.93           | 0.94        | 0.94        | 0.99        | 6%         |     |
| <b>Average</b>                 | <b>RF</b>  | <b>0.96</b>    | <b>0.99</b> | <b>0.99</b> | <b>0.97</b> | <b>1%</b>  |     |
|                                | <b>PER</b> | <b>0.66</b>    | <b>0.82</b> | <b>0.71</b> | <b>0.95</b> | <b>29%</b> |     |

**Three parameters are missing:**

| Missing Parameters                    | Model | Youden's Index | Precision | Specificity | Sensitivity | %FP | AUC  |
|---------------------------------------|-------|----------------|-----------|-------------|-------------|-----|------|
| <b>HR + ARTBPS + ARTBPM</b>           | RF    | 0.83           | 0.93      | 0.94        | 0.89        | 6%  | 0.97 |
|                                       | PER   | 0.04           | 0.51      | 0.1         | 0.93        | 90% |      |
| <b>HR + ARTBPS + CVP</b>              | RF    | 0.9            | 0.97      | 0.97        | 0.93        | 3%  | 0.99 |
|                                       | PER   | 0.06           | 0.51      | 0.13        | 0.94        | 87% |      |
| <b>HR + ARTBPS + PAPD</b>             | RF    | 0.9            | 0.97      | 0.97        | 0.93        | 3%  | 0.99 |
|                                       | PER   | 0              | 0.5       | 0.01        | 0.99        | 99% |      |
| <b>HR + ARTBPS + RR total</b>         | RF    | 0.9            | 0.97      | 0.97        | 0.93        | 3%  | 0.99 |
|                                       | PER   | 0.13           | 0.54      | 0.24        | 0.89        | 76% |      |
| <b>HR + ARTBPS + RR mandatory</b>     | RF    | 0.9            | 0.97      | 0.97        | 0.93        | 3%  | 0.99 |
|                                       | PER   | 0.12           | 0.53      | 0.23        | 0.89        | 77% |      |
| <b>HR + ARTBPM + CVP</b>              | RF    | 0.92           | 0.97      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.67           | 0.88      | 0.89        | 0.78        | 11% |      |
| <b>HR + ARTBPM + PAPD</b>             | RF    | 0.92           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.69           | 0.84      | 0.84        | 0.86        | 16% |      |
| <b>HR + ARTBPM + RR total</b>         | RF    | 0.93           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.69           | 0.84      | 0.84        | 0.86        | 16% |      |
| <b>HR + ARTBPM + RR mandatory</b>     | RF    | 0.92           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.69           | 0.84      | 0.84        | 0.86        | 16% |      |
| <b>HR + CVP + PAPD</b>                | RF    | 0.93           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.77           | 0.89      | 0.89        | 0.87        | 11% |      |
| <b>HR + CVP + RR total</b>            | RF    | 0.93           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.77           | 0.89      | 0.89        | 0.87        | 11% |      |
| <b>HR + CVP + RR mandatory</b>        | RF    | 0.94           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.77           | 0.89      | 0.89        | 0.87        | 11% |      |
| <b>HR + PAPD + RR total</b>           | RF    | 0.94           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.72           | 0.84      | 0.84        | 0.89        | 16% |      |
| <b>HR + PAPD + RR mandatory</b>       | RF    | 0.93           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.72           | 0.84      | 0.84        | 0.89        | 16% |      |
| <b>HR + RR total + RR mandatory</b>   | RF    | 0.94           | 0.98      | 0.98        | 0.95        | 2%  | 0.99 |
|                                       | PER   | 0.72           | 0.84      | 0.84        | 0.89        | 16% |      |
| <b>ARTBPS + ARTBPM + CVP</b>          | RF    | 0.89           | 0.96      | 0.97        | 0.92        | 3%  | 0.99 |
|                                       | PER   | 0.15           | 0.54      | 0.27        | 0.88        | 73% |      |
| <b>ARTBPS + ARTBPM + PAPD</b>         | RF    | 0.9            | 0.96      | 0.96        | 0.93        | 4%  | 0.99 |
|                                       | PER   | 0.01           | 0.5       | 0.01        | 1           | 99% |      |
| <b>ARTBPS + ARTBPM + RR total</b>     | RF    | 0.9            | 0.96      | 0.97        | 0.93        | 3%  | 0.99 |
|                                       | PER   | 0.2            | 0.55      | 0.22        | 0.97        | 78% |      |
| <b>ARTBPS + ARTBPM + RR mandatory</b> | RF    | 0.9            | 0.96      | 0.97        | 0.93        | 3%  | 0.99 |
|                                       | PER   | 0.2            | 0.55      | 0.22        | 0.97        | 78% |      |

| Missing Parameters                      | Model      | Youden's Index | Precision   | Specificity | Sensitivity | %FP        | AUC |
|---|------------|----------------|-------------|-------------|-------------|------------|-----|
| <b>ARTBPS + CVP + PAPD</b>              | RF         | 0.96           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.76           | 0.9         | 0.9         | 0.86        | 10%        |     |
| <b>ARTBPS + CVP + RR total</b>          | RF         | 0.97           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.26           | 0.57        | 0.27        | 1           | 73%        |     |
| <b>ARTBPS + CVP + RR mandatory</b>      | RF         | 0.97           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.24           | 0.57        | 0.29        | 0.96        | 71%        |     |
| <b>ARTBPS + PAPD + RR total</b>         | RF         | 0.97           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.01           | 0.5         | 0.01        | 1           | 99%        |     |
| <b>ARTBPS + PAPD + RR mandatory</b>     | RF         | 0.97           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.01           | 0.5         | 0.01        | 1           | 99%        |     |
| <b>ARTBPS + RR total + RR mandatory</b> | RF         | 0.97           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.22           | 0.56        | 0.22        | 1           | 78%        |     |
| <b>ARTBPM + CVP + PAPD</b>              | RF         | 0.97           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.9            | 1           | 1           | 0.9         | 0%         |     |
| <b>ARTBPM + CVP + RR total</b>          | RF         | 0.98           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.9            | 1           | 1           | 0.9         | 0%         |     |
| <b>ARTBPM + CVP + RR mandatory</b>      | RF         | 0.98           | 0.99        | 0.99        | 0.98        | 1%         | 1   |
|   | PER        | 0.9            | 1           | 1           | 0.9         | 0%         |     |
| <b>ARTBPM + PAPD + RR total</b>         | RF         | 0.98           | 0.99        | 0.99        | 0.99        | 1%         | 1   |
|   | PER        | 0.9            | 0.94        | 0.94        | 0.96        | 6%         |     |
| <b>ARTBPM + PAPD + RR mandatory</b>     | RF         | 0.98           | 0.99        | 0.99        | 0.99        | 1%         | 1   |
|   | PER        | 0.9            | 0.94        | 0.94        | 0.96        | 6%         |     |
| <b>ARTBPM + RR total + RR mandatory</b> | RF         | 0.98           | 0.99        | 0.99        | 0.99        | 1%         | 1   |
|   | PER        | 0.89           | 0.94        | 0.94        | 0.95        | 6%         |     |
| <b>CVP + PAPD + RR total</b>            | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1   |
|   | PER        | 0.99           | 1           | 1           | 0.99        | 0%         |     |
| <b>CVP + PAPD + RR mandatory</b>        | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1   |
|   | PER        | 0.99           | 1           | 1           | 0.99        | 0%         |     |
| <b>CVP + RR total + RR mandatory</b>    | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1   |
|   | PER        | 0.99           | 1           | 1           | 0.99        | 0%         |     |
| <b>PAPD + RR total + RR mandatory</b>   | RF         | 0.99           | 1           | 1           | 0.99        | 0%         | 1   |
|   | PER        | 0.93           | 0.94        | 0.94        | 0.99        | 6%         |     |
| <b>Average</b>                          | <b>RF</b>  | <b>0.94</b>    | <b>0.98</b> | <b>0.98</b> | <b>0.96</b> | <b>2%</b>  |     |
|   | <b>PER</b> | <b>0.54</b>    | <b>0.76</b> | <b>0.61</b> | <b>0.93</b> | <b>39%</b> |     |

# References

## Academic References

Aboukhalil, A., Nielsen, L., Saeed, M., Mark, R. G., & Clifford, G. D. (2008). Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *Journal of Biomedical Informatics*, 41 (3), 442-451.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, 308(6943), 1552.

Altman, D. G., & Bland, J. M. (1994). Statistics Notes: Diagnostic tests 2: predictive values. *British Medical Journal*, 309 (6947), 102.

Antink, C. H., Leonhardt, S., & Walter, M. (2016). Reducing false alarms in the ICU by quantifying self-similarity of multimodal biosignals. *Physiological Measurement*, 37 (8), 1233-1252.

Bishop, C. M. (2007). *Pattern recognition and machine learning*. New York: Springer.

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.

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

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30 (7), 1145-1159.

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

Cvach, M. (2012). Monitor alarm fatigue: an integrative review. *Biomedical Instrumentation & Technology*, 46 (4), 268-277.

DeVita, M. A., Smith, G. B., Adam, S. K., Adams-Pizarro, I., Buist, M., Bellomo, R., Bonello, R., ... Winters, B. (2010). "Identifying the hospitalised patient in crisis"—A consensus conference on the afferent limb of Rapid Response Systems. *Resuscitation*, 81 (4), 375-382.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *Bmc Bioinformatics*, 7 (1), 3.

Drew, B. J., Califf, R. M., Funk, M., Kaufman, E. S., Krucoff, M. W., Laks, M. M., ... & Van Hare, G. F. (2004). Practice standards for electrocardiographic monitoring in hospital settings. *Circulation*, 110 (17), 2721-2746.

Drew, B. J., Harris, P., Schindler, D., Salas-Boni, R., Bai, Y., Tinoco, A., Ding, Q., ... Mammone, T. (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).

- Drews, F. A. (2008). Patient monitors in critical care: lessons for improvement. In *Advances in patient safety: new directions and alternative approaches* (Vol 3: Performance and Tools, pp. 294–306). Rockville: MD.
- ECRI Institute (2010). Top 10 health technology hazards for 2011. *Health Devices*, 39 (11), 404-416.
- ECRI Institute (2011). Top 10 health technology hazards for 2012. *Health Devices*, 40 (11), 358-373.
- ECRI Institute (2012). Top 10 health technology hazards for 2013. *Health Devices*, 41 (11), 342-365.
- ECRI Institute (2013). Top 10 health technology hazards for 2014. *Health Devices*, 42 (11), 354-380.
- ECRI Institute (2014). Top 10 health technology hazards for 2015. *Health Devices*.
- ECRI Institute (2015). Top 10 health technology hazards for 2016. *Health Devices*.
- ECRI Institute (2016). Top 10 health technology hazards for 2017. *Health Devices*.
- Edworthy, J. (1994). The design and implementation of non-verbal auditory warnings. *Applied Ergonomics*, 25 (4), 202-210.
- Eerikäinen, L. M., Vanschoren, J., Rooijakkers, M. J., Vullings, R., & Aarts, R. M. (2016). Reduction of false arrhythmia alarms using signal selection and machine learning. *Physiological measurement*, 37 (8), 1204-1216.
- Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.
- Gazarian, P. K. (2014). Nurses' response to frequency and types of electrocardiography alarms in a non-critical care setting: A descriptive study. *International Journal of Nursing Studies*, 51 (2), 190-197.
- Gescheider, G. A. (1997). *The effects of skin temperature on the detection and discrimination of tactile stimulation*. New York: Guildford Publications.
- Goldstein, B. A., Polley, E. C., & Briggs, F. B. (2011). Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, 10 (1)
- Green, D., & Swets, J. (1996). *Signal detection theory and psychophysics*. New York: Wiley.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 (1), 29-36.
- Higham, P. A., & Arnold, M. M. (2007). Beyond reliability and validity: The role of metacognition in psychological testing. *New developments in psychological testing*, 139-162.
- Ho, T. K. (1995). Random decision forests, *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, August 1995.
- Imhoff, M., & Kuhls, S. (2006). Alarm algorithms in critical care monitoring. *Anesthesia and Analgesia*, 102 (5), 1525-1537.



Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 35.

Joint Commission on Accreditation of Healthcare Organizations (2002). Preventing ventilator-related deaths and injuries. *Sentinel Event Alert*, 25, 1-3.

Kam, P. C., Kam, A. C., & Thompson, J. F. (1994). Noise pollution in the anaesthetic and intensive care environment. *Anaesthesia*, 49 (11), 982-986.

Koski, E. M. J., Sukuvaara, T., Mäkitvirta, A., & Kari, A. (1994). A knowledge-based alarm system for monitoring cardiac operated patients-assessment of clinical performance. *International Journal of Clinical Monitoring and Computing*, 11 (2), 79-83.

Kumar, M., & M, T. (2006). Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest. *Ssrn Electronic Journal*.

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care and Pain*, 8 (6), 221-223.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2 (3), 18-22.

Little, R. J. A., & Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. New York, NY: John Wiley & Sons.

Lusted, L. B. (1971). Signal Detectability and Medical Decision-Making. *Science*, 171 (3977), 1217-1219.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, N.J: Lawrence Erlbaum Associates.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press.

Marino, P. L. (2013). *Marino's The ICU Book*. (2013). Wolters Kluwer.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology*, 50, 215-41.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Mohri, M., Talwalkar, A., & Rostamizadeh, A. (2012). *Foundations of machine learning*. Cambridge, Massachusetts: MIT Press.

Müller, B., Hasman, A., & Blom, J. A. (1997). Evaluation of automatically learned intelligent alarm systems. *Computer methods and programs in biomedicine*, 54(3), 209-226.

Novaes, M. A. F. P., Knobel, E., Bork, A. M., Pavao, O. F., Nogueira-Martins, L. A., & Ferraz, M. B. (1999). Stressors in ICU: perception of the patient, relatives and health care team. *Intensive Care Medicine*, 25, 1421-1426.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.

- Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, 4 (4), 171-212.
- Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *In ICML*, 98, 445-453.
- Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1 (1), 81-106.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufmann.
- Raghavan, V., Bollmann, P., & Jung, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *Acm Transactions on Information Systems (tois)*, 7 (3), 205-229.
- Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *Ibm Journal of Research and Development*, 3 (3), 210-229.
- Sendelbach, S., & Funk, M. (2013). Alarm Fatigue: A Patient Safety Concern. *Aacn Advanced Critical Care*, 24 (4), 378-386.
- Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A., & Horvath, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology : an Official Journal of the United States and Canadian Academy of Pathology, Inc*, 18 (4), 547-557.
- Singla, P., & Domingos, P. (2005). Discriminative Training of Markov Logic Networks. *Proceedings of the National Conference on Artificial Intelligence*, 20, 868-873.
- Siroky, D. S. (2009). Navigating Random Forests and related advances in algorithmic modeling. *Statistics Surveys*, 3 (0), 147-163.
- Sorkin, R. D. (1988). FORUM: Why are people turning off our alarms?. *The Journal of the Acoustical Society of America*, 84 (3), 1107-1108.
- Spackman, K. A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the sixth international workshop on Machine learning* (pp. 160-163). San Mateo: CA.
- Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62 (1), 77-89.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American*, 283 (4), 82-87.
- Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, 68, 301-340.
- Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61 (6), 401-409.
- Tsien, C. L. (2000). Event Discovery in Medical Time-series Data. *American Medical Informatics Association*, 7, 858-862.

Vesin, A., Azoulay, E., Ruckly, S., Vignoud, L., Rusinová, K., Benoit, D., Soares, M., ... Timsit, J. F. (2013). Reporting and handling missing values in clinical studies in intensive care units. *Intensive Care Medicine*, 39 (8), 1396-1404.

Wallis, L. (2010). Alarm Fatigue Linked to Patient's Death. *Ajn, American Journal of Nursing*, 110 (7), 16.

Ward, M. M., Pajevic, S., Dreyfuss, J., & Malley, J. D. (2006). Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests. *Arthritis Care & Research*, 55 (1), 74-80.

Willich, S. N., Wegscheider, K., Stallmann, M., & Keil, T. (2005). Noise burden and the risk of myocardial infarction. *European Heart Journal*, 27 (3), 276-282.

Xie, H., Kang, J., & Mills, G. H. (2009). Clinical review: The impact of noise on patients' sleep and the effectiveness of noise reduction strategies in intensive care units. *Critical Care*, 13 (2), 208.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3 (1), 32-35.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115 (5), 654-657.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39 (4), 561-77.

## תקציר

**רקע:** מערכות ניטור רפואיות מהוות חלק בלתי נפרד מהסביבה הרפואית של המטופל ונועדו להסב את תשומת ליבם של הרופאים לשינויים חריגים במצבו. מערכות אלה הינן בעלות רמת רגישות גבוהה במיוחד על מנת שלא יפספסו אף שינוי במדדים הפיסיולוגיים של המטופל. בעקבות זאת, הספציפיות של מערכות אלה הופחתה לטובת רגישות גבוהה יותר. מחקרים רבים חושפים כי שיעור אזהקות השווא (FAR) של מערכות הניטור נע בין 72% ל-99%. שיעור גבוהה זה גורם לצוות הרפואי להנמיך את עוצמת האזהקה, לכבות ואף להתעלם ממנה. מקרים כגון אלה הובילו למותם של מטופלים רבים.

**שיטות:** במחקר זה, אנו משתמשים בשיטה המבוססת על סט חוקים שפותחו על ידי רופא מומחה להפחתת שיעור אזהקות השווא ונבדקו על נתונים שנאספו ביחידת טיפול הנמרץ של המרכז הרפואי בתל אביב (TAMC). שיטה זו מצליחה נתונים ממספר חיישנים של מערכת הניטור לשם אימות המידע וכאשר פרמטר אחד או יותר עוברים את הסף המוגדר לדקה אחת או יותר, תשמע אזהקה. שיטה זו נשענת על נתונים ממספר חיישנים שונים שנדגמו באותה דקה; עם זאת, לעתים קרובות לא כל החיישנים זמינים. היעדר הפרמטרים מביא לביצועים ירודים ולעלייה בשיעור אזהקות השווא. על מנת להתגבר על בעיה זו, השתמשנו בשיטה המבוססת על למידת מכונה (ML) אשר בה מתבצעת הכללה מהפרט אל הכלל; וכך, מצליחה לנבא תחזיות על נתונים שלא ראתה קודם לכן ומאפשרת להתגבר על נתונים חסרים. לשם כך, אימנו מודל *random forest* על נתונים המכילים אחד עד שלושה פרמטרים חסרים, כאשר ההגדרות מבוססות המומחה FER (ללא פרמטרים חסרים) אשר פותחו במסגרת מחקר זה שימשו כאמת המוחלטת. תוצאות הסיווג של מודל *random forest* עם הפרמטרים החסרים הושו לתוצאות מודל ההגדרות מבוססות המומחה החלקי PER, כאשר בו כללים המכילים את הפרמטרים החסרים הוסרו בהתאמה. הערכת המודלים נעשתה באמצעות precision, sensitivity, specificity ו-Youden's statistic (index) אשר מבוסס על שני המדדים הראשונים.

**תוצאות:** במבחן הראשון כל האזהקות (Bradycardia, Bradycardia hypotension, Hypovolemia, Tachycardia, Tachycardia hypotension, Obstructive shock, Left ventricular (LV) shock) נבדקו יחד, כאשר בבדיקה השנייה כל אחת נבדקה בנפרד. על פי תוצאות הבדיקה הראשונה, החוסר בנתוני קצב הלב (HR), לחץ הדם העורקי סיסטולי (ARTBPS), לחץ הדם העורקי הממוצע (ARTBPM) ולחץ הדם העורקי ריאתי (PAPD) הוביל לביצועים הגרועים ביותר של RF ו-Youden precision נעו בין 0.94 ל-0.99, 0.98 ו-0.99 עבור RF בהשוואה ל-PER אשר נע בין 0.54 לבין 1, 0.76 ו-1 בהתאמה. במבחן השני, מדד Youden ומדד precision נעו בין 0.43 ל-1, 0.65 ו-1 עבור RF ו-0.25 ו-1, 0.58 ו-1 עבור PER.

**מסקנות:** גישה זו של למידת מכונה הוכיחה כי יכולת קבלת החלטותיה לאחר למידה המבוססת על ידע של רופא מומחה עם ניסיון רב שנים הינה ברמה השווה לזו של המומחה במידה ואין מחסור בנתונים. עם זאת, בעוד מודל RF מגיע לדיוק גבוהה ו-FAR נמוך, הביצועים של מודל PER הולכים ופוחתים עם מספר החיישנים החסרים. לדוגמה, החלטות שהתקבלו באמצעות מודל PER הציגו FAR של 6%, 11%, ו-23% כאשר אחד, שניים, או שלושה חיישנים חסרים, בהתאמה, ואילו אלה של RF הובילו ל-FAR יציב שנע בין 2-3%. אנו מייחסים את הצלחת ה-RF ליכולתו להצלבת המידע מהחיישנים הזמינים המפצים על החסר, ומראים כי עיבוד המידע הנוכחי באמצעות המוח האנושי יעיל פחות מהגישה המוצעת של ML.

**מילות מפתח:** יחידת טיפול נמרץ, אזעקות שווא, Random Forest, Machine Learning



אוניברסיטת בן-גוריון בנגב  
הפקולטה למדעי ההנדסה  
המחלקה להנדסת תעשייה וניהול

## **שיטה לצמצום אזעקות שווא וההשפעה של נתונים חסרים ביחידה לטיפול נמרץ על ידי יישום שיטות של למידת מכונה והצלבת מידע ממספר חיישנים**

מאת: גל חבר

בהנחיית: דר' יובל ביתן

חיבור זה מהווה חלק מהדרישות לקבלת תואר מגיסטר בהנדסה

תאריך: 26-10-2017

חתימת המחבר: גל חבר ...

תאריך: 26-10-2017

אישור המנחה: דר' יובל ביתן

תאריך: 26-10-2017

אישור יו"ר ועדת תואר שני מחלקתית: פר' ישראל פרמט

אוקטובר 2017

חשון תשע"ח



אוניברסיטת בן-גוריון בנגב  
הפקולטה למדעי ההנדסה  
המחלקה להנדסת תעשייה וניהול

# **שיטה לצמצום אזעקות שווא וההשפעה של נתונים חסרים ביחידה לטיפול נמרץ על ידי יישום שיטות של למידת מכונה והצלבת מידע ממספר חיישנים**

מאת: גל חבר

בהנחיית: דר' יובל ביתן

חיבור זה מהווה חלק מהדרישות לקבלת תואר מגיסטר בהנדסה

אוקטובר 2017

חשון תשע"ח