# Machine learning for automated discovery of clinically unstable episodes in paediatric intensive care patients with congenital heart disease: Aberration Detection

R. Zoodsma (6355404)

Supervisor: J. Nijman, MD PhD

Department of Pediatric Intensive Care, Universitary Medical Center Utrecht / Wilhelmina Children Hospital Utrecht, The Netherlands

07-01-2022 – 03-04-2022

# Abstract

## Objectives

With the ongoing shift from patient wards to single-person rooms, monitoring of the patient's condition (outside the room) can become challenging, especially when combined with the complex physiology of perioperative Congenital Heart Defects (CHD). Machine learning (ML) may support the medical team through automated detection of clinical deterioration. In this proof-of-concept study, it was aimed to develop a dual-approach aberration-detection algorithm in Paediatric Intensive Care Unit (PICU) patients with CHD.

## Methods

Data of four vital parameters and cerebral rSO2 of neonates with complex CHD admitted to the University Medical Centre Utrecht, The Netherlands between 2002 and 2018 were used for training. This dataset was integrated into an algorithm, designed to detect both population-specific abnormal parameter combinations using Support Vector Machine learning, as well as significant patient baseline deviations. The algorithm was applied on test data from new patients and subsequently visualized. The aberration-detection was evaluated by a single expert in the field.

## Results

A respective 4600h and 229h in 78 and 10 neonates were used as training and test dataset. Four examples of patient-specific aberration-detection visualizations are provided in appendix B. Overall, the algorithm provided accurate detection in 89,8% of stable- and 71,3% of unstable episodes. Twenty-nine out of 101 unstable periods were missed in testing.

## Conclusions

ML can be used to automatically classify big PICU time-series datasets, although accuracy should be improved and prospectively evaluated. ML-based classification algorithms in the PICU setting may, eventually, provide an addition to conventional monitoring and enhance the processing of big datasets for research.

# I - Introduction

Globally, congenital heart defects (CHD) form the most common of congenital anomalies in newborn. Though exact prevalence at birth is unknown due to a variety of factors, estimations mention 8-10 CHD per 1000 live births(1,2). The clinical presentation of CHD infants can vary tremendously from asymptomatic (*e.g., small atrial septum defects)* to life-threatening directly after birth *(e.g., hypoplastic left heart syndrome)*. Depending on the diagnosis, ±25% of CHD infants require cardiac surgery at some point in their lives (1–4). With a tremendous growth in overall survival rate observed over the last decades, research focus over the last 5-10 years has shifted from increasing survival rate to decreasing morbidity(2,4–6).

Especially during the peri-operative period in the more complex CHD, infants are vulnerable to a wide range of complications such as neurological defects, heart failure or necrotizing enterocolitis. These adverse events may have a severe impact on quality of later life, as most infants born in the 21$^{st}$ century reach adulthood(2,5,6). In the prevention of such adverse events, intensive monitoring of vital functions are key: several studies have recently shown distinct trends in various parameters to be associated with adverse outcomes(7–9). Current Dutch PICU-units determine the need for frequent evaluation of vital functions and, if required, timely intervention by use of the Pediatric Early Warning Score (PEWS). The PEWS award points to parameters exceeding age-specific cutoff values, where a higher cumulative score may be seen as a worse clinical condition requiring more frequent monitoring(10,11).

In a PICU-setting with specialistic CHD-care however, value ranges for vital functions deemed 'normal' may differ severely depending on underlying etiology(1,3). These differing normal value ranges make the subtle changes in vital functions, which can precede clinical instability, increasingly difficult to pay attention to. Considering different levels of training, combined an ongoing shift from PICU-wards to single-person rooms, the monitoring of the patient's condition outside the room can become quite challenging. For this task of detecting novel episodes of patient deterioration, also known as Aberration Detection, machine learning (ML) may prove useful.

Over the last decade, many ML models in different populations have been designed to support the medical team in the automated prediction of future clinical endpoints, with varying results(8,9,12–14). Most models considered a ML approach to Early Warning Systems (EWS), designed to tackle one of the EWS's greatest pitfalls: the underlying assumption of independence between vital parameters(9,13). In a recent systematic review concerning ML in child-medical settings, only 2% *(n=6)* studies were performed in an Intensive Care setting. None of the examined studies however considered infants with CHD, highlighting the current gap in research of this fragile population (14).

In this proof-of-concept paper, we present a dual-approach aberration-detection algorithm (AD-Algorithm) designed to detect novel deterioration in PICU patients with CHD. The AD-Algorithm utilizes both a Support Vector Machine (SVM) to distinguish clinically stable- from unstable combinations of vital parameters whilst taking underlying correlation into account, as well as significant patient-specific baseline deviations in order to detect possible patient deterioration. The study design will be explained in chapter II, where chapter III focuses on the algorithmic structure. Results can be found in chapter IV; chapter V is used for discussion and states both room for improvement as well as further research. Chapter VI concludes the concept.

# II – Study design

## Patient selection

Retrospectively, 113 CHD infants admitted to the PICU of our hospital between January 2002 and December 2018 were included in a training-dataset. Patients were included based on availability of reliable, time-matched monitor data.

Out of the 113 initially included patients, 35 (*31%*) were excluded (figure 1) for one of following five reasons:

**1)** Missing or invalid patient ID (*n=14)*,
**2)** Less than 12 hours of data was available (*n=11*),
**3)** No documented CHD (*n=5*),
**4)** Birth weight under 2000g (*n=3*), or
**5)** Age > 1y at time of first measurement (*n=2*).

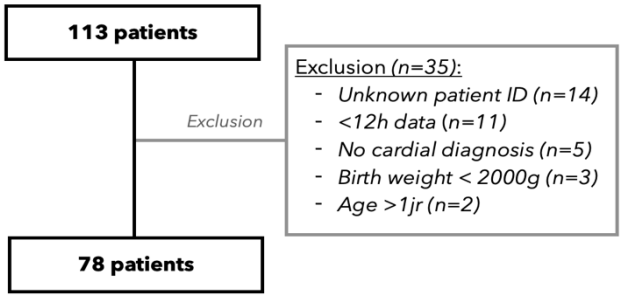



*Figure 1: Patient flowchart depicting exclusion criteria.*

## Data collection- & preprocessing

Time-matched monitor-data was extracted for each patient using a sample-interval of one second for the following six parameters:

- Arterial oxygen saturation (SpO2)
- Regional brain oxygen saturation (rSo2)
- Mean Invasive blood pressure ($IBP_{mean}$)
- Respiratory rate (RR)
- End-tidal CO2 (EtCO2)
- Heart rate (HR)

No imputation was performed to account for missing values. EtCO2 was used to determine whether patients were actively being ventilated. We consider a patient at time $\boldsymbol{t}$ on active ventilation if the corresponding EtCO2 is greater than zero. As EtCO2 is always registered when on active ventilation, the value may be absent yet never missing. We therefore consider both an absent- as well as null value of EtCO2 to represent a patient not on active ventilation at the specific time.

RR values were observed to differ greatly within minutes, where we argue the overall RR be better examined when looking at either an upward- or downward trend rather than absolute values. A 300s moving average of respiratory rate preceding each time $\boldsymbol{t}$ was consequently implemented.

rSo2 is measured using two cerebral probes attached to the left- and right forehead, approximating regional brain saturation in their respective frontal lobe. In certain cases, for example when the infant's head is simply not large enough for both probes, a single probe is used. When a single probe is functional, rSo2 is equal to that value. If both probes are transmitting however, the mean value is used for calculations.

## Group division

As stated before, value ranges considered normal may vary greatly depending on underlying etiology. Globally, these normal value ranges may be split into two groups depending on the presence- or absence of a cyanotic CHD, where poorly-oxygenated- is mixed with highly oxygenated blood prior to being pumped towards peripheral tissue. As a result, arterial oxygen saturation (SpO2) is significantly lower in patients with a cyanotic CHD when compared to the SpO2 of non-cyanotic CHD infants(3,4).

To overcome these varying values deemed normal, patients were divided over two groups based on average SpO2 during admission: $SpO2_{avg}$ <90 versus $SpO2_{avg}$ >90. Average SpO2, instead of the presence- or absence of cyanotic CHD, was chosen following several cases combining a cyanotic CHD with a high pulmonary flow: these specific cases were deemed a better fit in the other group. Subsequently, division was performed based on average SpO2 during admission, instead of the absence- or presence of a cyanotic CHD.

Out of the included 78 patients, 26 (*33.3%*) were included in group *Alpha*, with an average SpO2 < 90. Group *Beta* consisted of 52 patients (*66.7%*) with an average SpO2 > 90. Characteristics of both subgroups can be found in table 1 below.

**Table 1: Baseline characteristics of group Alpha ($SpO2_{avg}$ < 90) and Beta ($SpO2_{avg}$ > 90).**

| | $SpO2_{avg}$ < 90 | $SpO2_{avg}$ > 90 | |
|---|---|---|---|
| | *n = 26 (33,3%)* | *n = 52 (66,7%)* | |
| **Group characteristics** | | | |
| **Male gender** *(%)* | 21 (*80,8%)* | 35 (*67,3%)* | *p = 0,20* |
| **Birth weight** *(kg)* | 3,4 ±*0,6* | 3.3 ±*0,5* | *p = 0,54* |
| **Age at t=0** *(days)* | 7,0 (*8,7)* | 9,0 (*12,3)* | *p = 0,53* |
| **Available data** *(hours)* | 63,1 (*52,5)* | 44,0 (*37,8)* | ***p < 0,01*** |
| **Parameter mean values** | | | |
| **Heart rate** *(beats/min)* | 156,8 ±*8,4* | 144,4 ± *18,6* | ***p < 0,01*** |
| **Resp. rate** *(breaths/min)* | 34,4 ±*8,3* | 34,8 ±*8,7* | *p = 0,85* |
| **SpO2** *(%)* | 75,3 ± *7,7* | 95,6 ±*3,2* | ***p < 0,01*** |
| **rSo2** *(%)* | 55,7 ± *10,1* | 71,7 ± *12,0* | ***p < 0,01*** |
| **$IBP_{mean}$** *(mmhg)* | 52,5 ±*8,0* | 54,0 ±*9,8* | *p = 0,50* |

*Data are depicted as counts (percentages), mean ± standard deviation or median (interquartile range). P-value was calculated using the two-tailed, heteroscedastic independent sample T-test.*

# III – Algorithm structure

**Overall structure**

The overall target of the AD-algorithm is to detect episodes of clinical (in)stability in PICU-patients with CHD. The subgroups *Alpha* & *Beta*, as stated in chapter II, differed significantly in both mean parameter values (*HR, SpO2 & rSo2)*, as well as underlying correlations by such extent that we argue a single algorithm would fail to accurately represent both groups at the same time. Therefore, two algorithms were created taking the above significant variations into account.
Overall setup remains similar (figure 2).

---

The conceptual foundation of the AD-algorithm relies on three separate models, each with their own focus: *A)* combined parameter (in)stability, *B)* significant baseline deviation and *C)* sensorial dysfunction. Together, the three models decide at each moment during admission whether the patient is either stable- or unstable.

Model A may be compared to a Paediatric Early Warning Score (PEWS). The PEWS scores each parameter separately to form a combined score which, when exceeding a pre-defined cutoff value, may indicate patient instability thus recommending care to be escalated to a higher level. Where the PEWS however assumes independence of corresponding parameters, model A utilizes Support Vector machine learning (SVM) to analyze the *combination* of parameters while taking underlying correlation into account(9,13). The SVM scores each entry in time separately to either be stable- or otherwise.

Where model A treats each entry as independent from another, model B was designed to detect events of significant parameter deviations *over time*. Current vital functions are compared to a unique, patient-specific baseline calculated using aggregate parameter values during admission. Detected significant deviations could however also be related to clinical improvement, rather than instability. By analyzing the *trend* of the deviation, improvement (*trend movement towards subgroup mean*) is distinguished from instability (*trend movement away from subgroup mean*).

Model C relates on sensorial dysfunction. Predefined, static cutoff values have been considered for $IBP_{mean}$, SpO2, RsO2 and RR. At times when these predefined conditions are met, we consider either equipment to not function properly or unrealistic values to be transmitted. An entry in time will consequently be treated as dysfunctional.

Model A & B both label data to be either stable or instable upon different grounds, which can result in conflicts in labeling. An entry in time is considered 'stable' if both models agree the specific time to be so. If either model determines the current status of the patient be unstable, we consider the corresponding entry to be unstable. Regardless of A and B, if model C determines any sensor to be dysfunctional, the entry will be labeled as such.

To prevent false-positive labeling of a short 'burst' of instability due to various reasons, we follow Clifton et al(9) in exclusively considering a period 'unstable' when labeled to be so in at least 80% of any 5-minute frame.

Models were built using RStudio 1.4 and will be discussed in detail below. Full access to the code used in constructing each model can be requested through the author. Visualization of aberration detection in four test patients are enclosed in Appendix A.

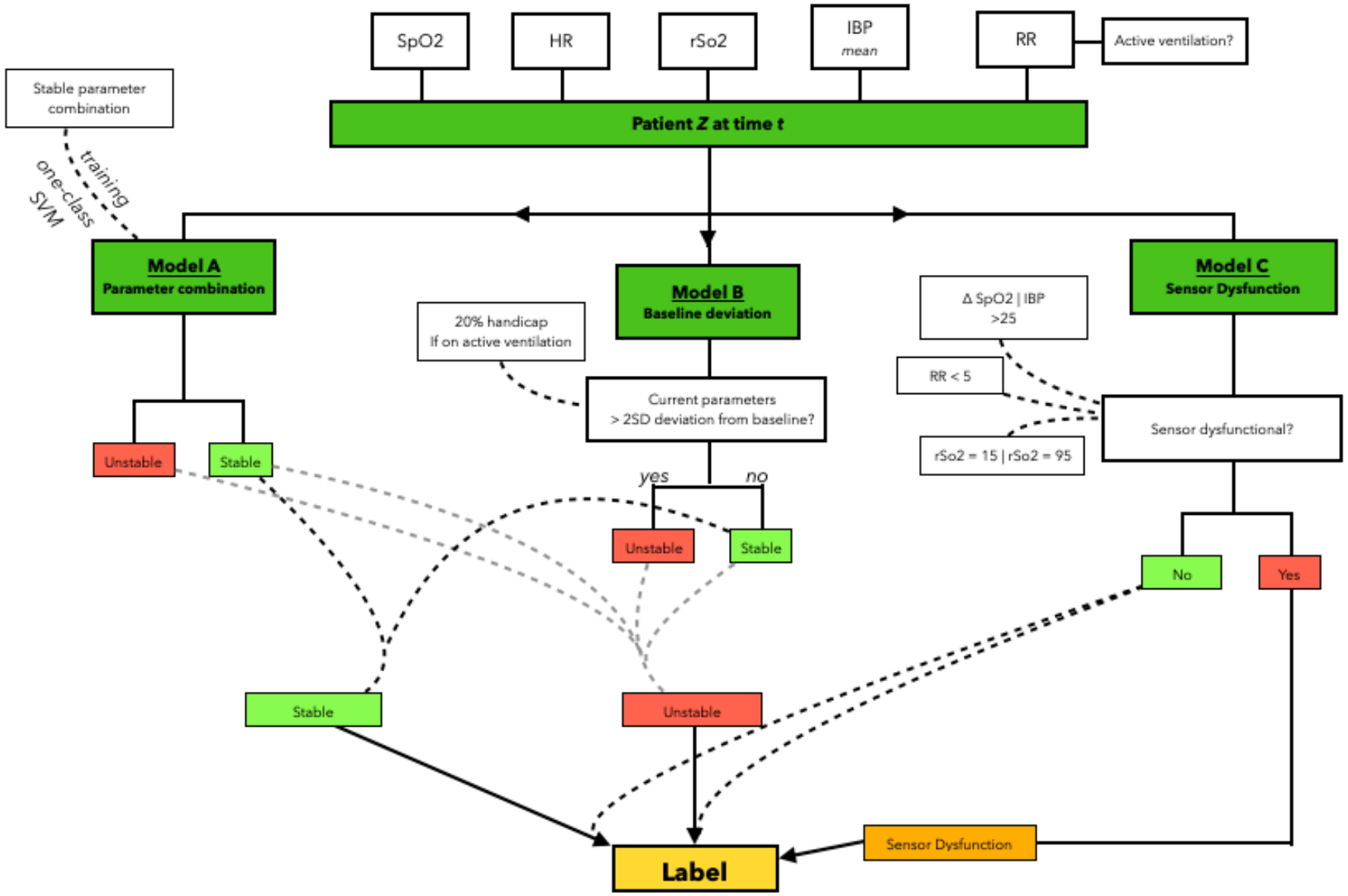



*Figure 2: Flowchart representing the labeling process using the three shown models.*

Model A – Combined parameter (in)stability

Each parameter was normalized with respect to their own, group-dependent, mean and variance. For any entry ***y***, i.e., vector of five parameters at time ***t***, the corresponding Mahalanobis distance (M-distance) was calculated. The M-distance quantifies the combined distance of an *n*-dimensional vector to their respective means, taking inter-dimensional correlation into account(15). Parameter correlations were calculated using the Pearson method (*appendix B)*.

Colleagues Clifton et al(9), for a similar purpose yet different population and setting, made use of a Kernel Density Estimation (KDE) in dividing stable- from unstable parameter combinations. We follow their proposed 80%-division cutoff value. However, where Clifton et al(9) deem the 20% of their kernels with the farthest distance to the origin to be 'unstable', we consider the parameter combinations 'stable' if their respective M-distance is lower than our 80$^{th}$ percentile cutoff value. Parameters and their respective M-distances in the top 20$^{th}$ percentile were considered 'unstable'.

A random 80 percent of parameters deemed 'stable' were used as training-dataset for a One-class Support vector machine (SVM). In training the SVM, a square-exponential Kernel was used. The degree by which datapoints are allowed to be misclassified in training, also known as parameter *'Mu'*, has been set to 5% in order to prevent overfitting the SVM to the training-dataset.

When predicting a previously unseen vector of parameters ***y*** at time ***t*** by the SVM, the vector can be predicted either as an 'outlier' (i.e., most likely *non*-resemblant to the training-dataset) or an 'inlier' (most likely resembling the training-dataset). Following a supposedly 'stable' combination of parameters in training the SVM, an in- or outlier will respectively be considered 'stable and 'unstable'.

Besides above prediction, model A also considers singular parameters exceeding a static cutoff value to be related to patient instability. Any heart rate above 200 beats/minute, respiratory rate above 70breaths/min or a mean invasive blood pressure under 30mm Hg were considered to be unstable, regardless of other parameter values(figure 3).

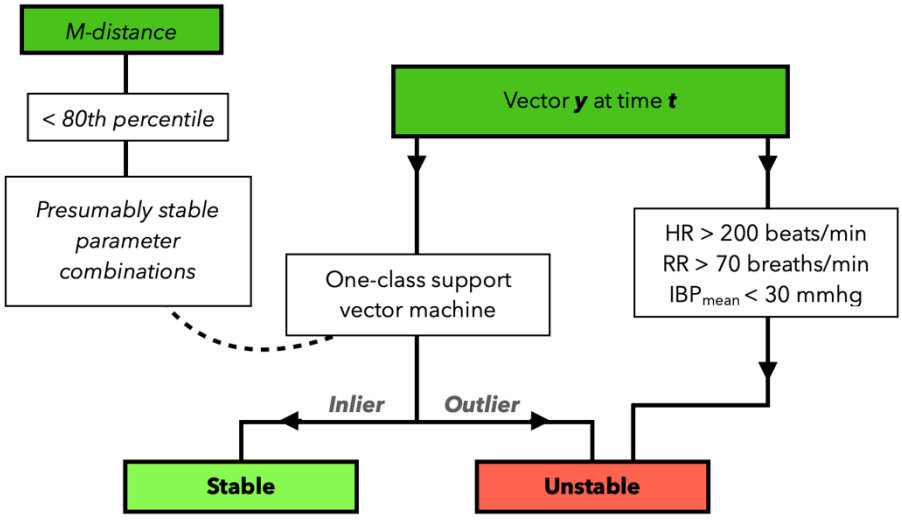



*Figure 3: Flowchart of model A. Vector y represents a set of parameters at time t.*

## Model B – Significant baseline deviation

Besides model A determining each entry in time to be stable- or otherwise, model B was implemented to rather analyze the *course* of parameters during admission instead of treating each point in time to be independent from previous measurements. For each patient, M-distance is compared to their unique baseline and subsequently analyzed. Variation of M-distance beyond double the standard deviation could either depict clinical instability- or improvement, model B considers both situations (figure 4).

The model B-design relies mainly on three calculations:

1) ***B*** -- Baseline M-distance
   The patient-specific baseline value was determined for each time *t* as the median of M-distances between *t=0* and the current time ***t***.
2) ***Z*** -- M-distance smoothed over 300s
   The current M-distance at time ***t*** was smoothed using a preceding 300s-moving average. Prior to smoothing, M-distance was multiplied by a factor 1.2 if on active ventilation at time ***t***. We argue patients on active ventilation to show a decreased variation in overall physiological state due to, among others, the sedated status as well as an iatrogenic fixed respiratory rate. Therefore, in this controlled setting a 'penalty' of 20% is introduced to account for the consequent reduction in vital function variation.
3) ***Z – B*** -- Standard deviation:
   Using the previously generated dataset of presumably stable parameter combinations used in training model A, the difference in M-distance ***Z - B*** at any time ***t*** is calculated by subtracting calculation 2 (*M-distance at time **t***) from calculation 1 (*Baseline M-distance*). The subsequent standard deviation (SD) of ***Z - B*** is calculated.

The calculated value of formula 3 can either be positive, where $Z - B > 0$ or negative, where $Z - B < 0$.

Scenario 1: $Z - B > 0$

As M-distance is calculated using normalized parameter values, an M-distance of 0 can be considered as all parameters precisely bearing the value of each respective mean. If the difference ***Z – B*** is a positive number, the current M-distance (as proxy for patient condition) lies farther away from the parameter means after baseline correction. We consider this scenario to be related to clinical deterioration, as parameters significantly drift away from both their respective means as well as the patient-specific baseline value. A subsequent label 'unstable' will be assigned.

Scenario 2: $Z - B < 0$

Following the argumentation at scenario one, current scenario is considered to be related to clinical improvement as vital parameters approach their respective means. The entry in time will be assigned the corresponding 'stable' label.

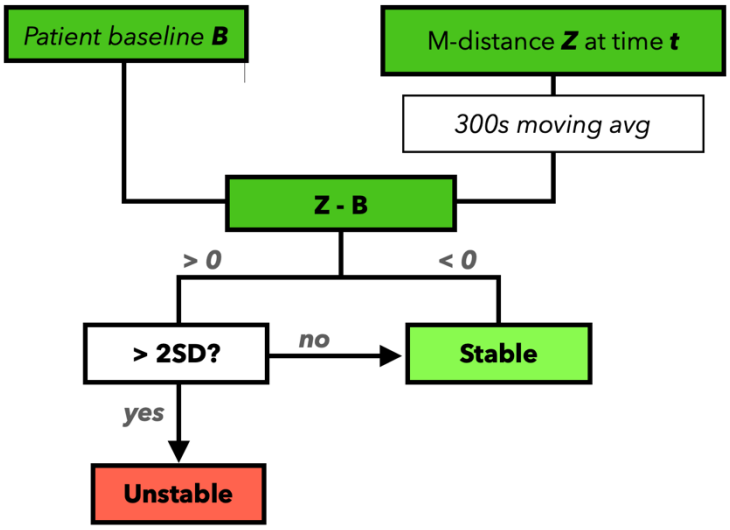



*Figure 4: Flowchart of model B.*

## Model C – Sensorial dysfunction

Model C was designed to detect sensorial artefacts, in order to prevent wrongfully considering an entry in time to be 'unstable' due to these sensor artefacts. The following parameter values were considered sensorial artefacts (*figure 5*):

- *$IBP_{mean}$ | SpO2*

At times when arterial catheters are used for blood draw, an immediate spike in $IBP_{mean}$ can be observed. Similar spikes in SpO2 (non-related to blood draw) may be seen. Therefore, a sensor artefact regarding $IBP_{mean}$ and SpO2 was defined as a difference with the preceding measurement of more than 25 points on their respective scale.

- *rSo2*

The rSo2-monitor depicts a value between 15 and 95. However, a sensorial error due to probe malfunction may result in a static high- or low value. Therefore, rSo2-values of either minimal value (15) or maximal value (95) are considered as probe malfunction.

- *Respiratory rate*

Upon active ventilation support, false RR monitor-values may at times be observed as negative- or close to zero. Therefore, any RR value upon active ventilation lower than five breaths/min is considered to be due to sensor malfunction.

- *Heart rate*

Heart-rate sensorial dysfunction can me quite tricky to model, as there is little difference between, for example, cardiac arrest and sensorial dysfunction: both could consequently drop to zero. As at this point in the study no reliable distinction between the two situations can be established, no heart-rate sensor dysfunction was modeled.

Upon artefact detection, both 60 seconds before- as well as 60 seconds after the last detection will be considered dysfunctional for two reasons. Firstly, monitor-data in the minute surrounding artefact detection may be unreliable due to, for example, patient movement. Secondly, most sensor artefacts were observed to be related to arterial blood draw, effectively reducing clinical usefulness of the algorithm as a member of the medical team is physically present in the room.

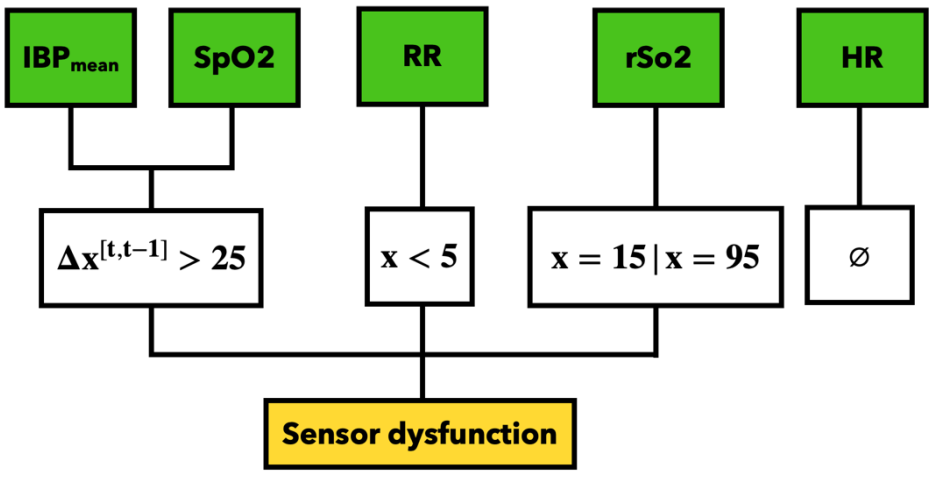



*Figure 5: Flowchart of model C. X represents the corresponding parameter value at any time **t**.*

# IV – Results

## Algorithm effectiveness evaluation

Evaluation of ML algorithms can be challenging. Different methods exist, where best practice is considered as independent labeling- and evaluation by experts in the field. Due to, specifically in this case, both poor reporting of clinical (in)stability in medical records (*medical notes are not written for the purpose of retrospective study analysis)* as well as the large amounts of available data *(over 4600h)*, this process can be extremely time consuming. A different method of determining model accuracy was therefore used, where effectiveness was evaluated through expert-analysis of algorithmic labeling, focusing on both the time-percentual correctness, as well as the number of unstable episodes missed.

Effectiveness was determined by a single expert in paediatric intensive care. Accurate labeling was defined as both proposed algorithm- and the expert agreeing on the matter, where inaccurate labeling was considered at times the algorithm did not match the expert's opinion. Both the number of (in)correctly labeled episodes, as well as the amount of time per (in)correct episode were noted. It is however important to note that, due to a disproportionate ratio of stable vs unstable episode duration, accuracy may be skewed towards a higher- or lower percentage. In machine learning, this phenomenon is known as *the accuracy paradox*.

For example, if labeling 10 hours of data resulted in 10 missed 6-minute unstable episodes, the model accuracy would add up to 90%, yet missing a significant amount of 10 episodes. It is therefore important to determine model effectiveness on both correctly labeled number of episodes, as well as the percentual time influence. To slightly account for *the accuracy paradox*, episodes were counted with a maximum duration of two hours (*e.g., a 10-hour stable period will consequently be counted as 5 episodes).* Sensorial dysfunctions were singularly analyzed in number of episodes right- or wrongfully labeled.

## Results

A total of 109 & 120 hours of data across both subgroups ($SpO2_{avg}$ < 90 & $SpO2_{avg}$ > 90 respectively) were assessed by a single expert in the field (figure 6).

In the subgroup regarding $SpO2_{avg}$ < 90, a total of 59 stable episodes occurred where 52 (*88.1%)* were correctly labeled. The 59 episodes lasted 92,5 hours, where 88,25 (*95.4%)* hours were correctly analyzed. Unstable episodes occurred 47 times for a total of 14,75 hours. Thirty-two episodes (*68.1%*) were correctly labeled, adding up to 8,0 hours (*54.2%*). A single unstable episode (*3,1%)* was correctly detected, yet algorithmic labeling did not cover the full length of this episode.

In the subgroup regarding $SpO2_{avg}$ > 90, stable episodes occurred 58 times out of which 53 (*91.4%)* were correctly labeled. Stable episodes lasted a total of 91,4 hours, where 87.7% (*80,7 hours*) were correctly labeled. Across 54 unstable episodes adding up to 25,8 hours, 21,1 hours (*81.9%*) over 40 episodes (*74.1%*) were labeled accordingly. Out of the 40 correctly detected unstable episodes, four (*10,0%*) were only partially correct.

Considering both groups, 107 (*89,8%)* of the 117 stable episodes were correctly labeled. Unstable episodes were rightfully labeled in 72 *(71,3%)* out of the 101 observed episodes. Sensor dysfunction occurred a total of 144 times, of which 136 (*94,5%)* were accurately labeled. Eight (*5,5%*) sensorial episodes were wrongfully detected where no dysfunctions were missed.

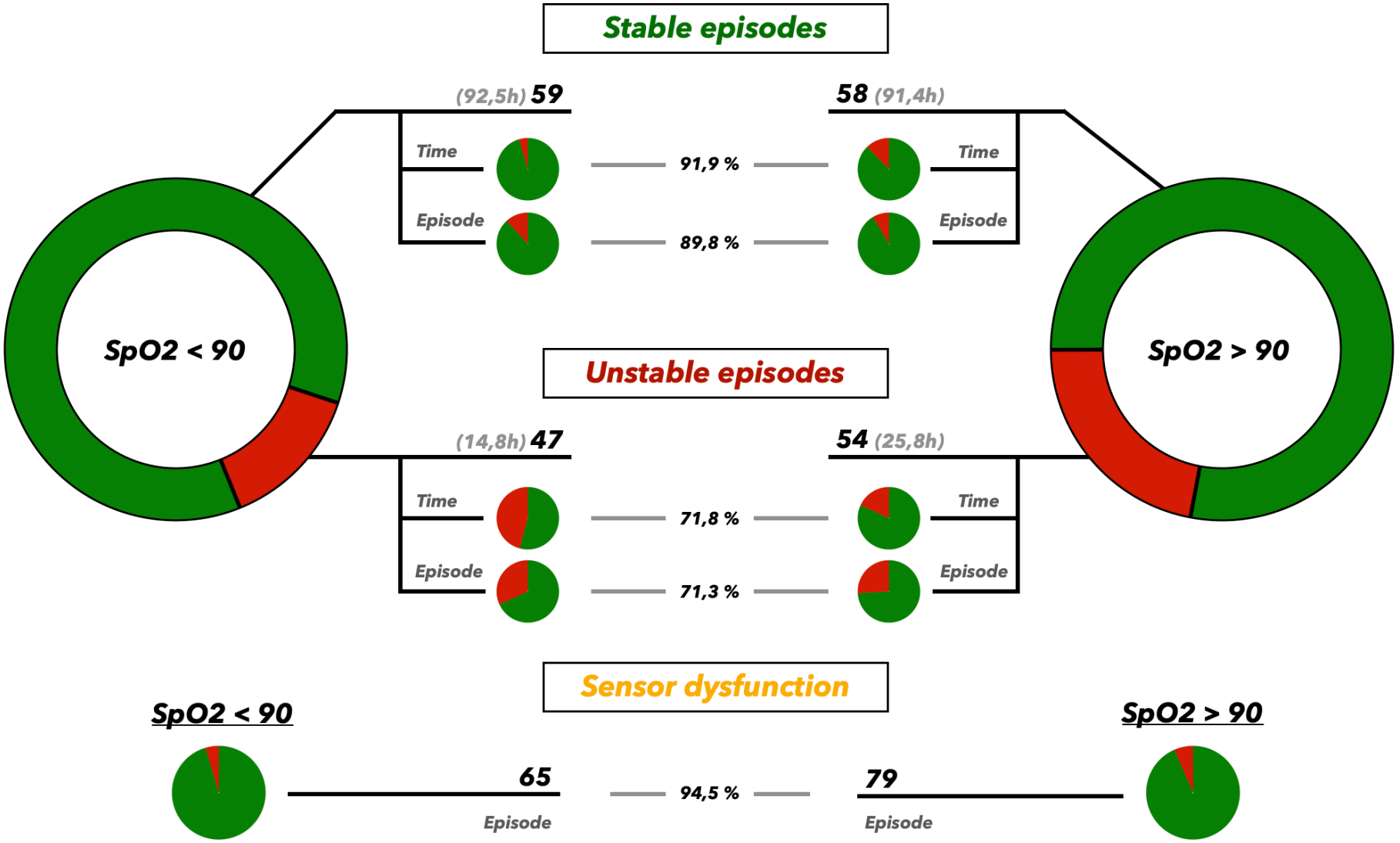



*Figure 6: Flowchart depicting automated labeling results using the AD-algorithm.*

# V – Discussion

In this proof-of-concept paper, a dual-approach aberration detection algorithm in PICU patients with CHD was developed. Overall, the algorithm provided accurate detection of respectively 89,8% and 71,3% of observed stable- and unstable episodes. Out of a total of 101 confirmed unstable episodes with varying length, 29 episodes (*28,7%*) were missed. Sensorial artefacts were accurately detected in 94,5% out of 144 detections.

Upon visual analysis of evaluated detections, several observations were made.

Firstly, more unstable episodes were wrongfully considered to be stable at a later stage of admission rather than at the start of measurement. This reducing sensitivity over time may be related to the aggregate M-distance used in calculating the unique baseline. For example, consider a patient clinically improving over the last day, yet relapsing at time ***t***. If relapse occurs a later stage in time, the resulting decrease in M-distance due to clinical improvement has not (yet) lowered the aggregate baseline far enough for the relapse at time ***t*** to be successfully detected. If the same situation however occurs at an earlier stage of measurement, the percentual influence on M-distance is much greater, effectively lowering the baseline and successfully detecting the relapse at time ***t***. This phenomenon may be overcome by considering both significant deviations from aggregate M-distance baseline, as well as deviations regarding a baseline calculated over a shorter period of time to be clinically significant.

Secondly, faulty sensor dysfunction occurred in 75% *(n=6)* of observed instances due to respiratory rate falling below 5 breaths per minute, whilst other parameters remained stable. Sensor errors regarding monitor-determined respiratory rate occurred only at times when patients were actively ventilated. The monitor-determined RR, as used in this study, is a derivation from thoracic movement registered through electrocardiography (ECG). However, the data directly from the ventilator itself may provide much more (reliable) information regarding assessment of respiratory capacity besides the number of 'breaths'/min. In a future stage of development, the assessment of respiratory capacity at times of active ventilation may be more accurately estimated using different available (ventilator-derived) parameters. Additionally, this would likely result in a reduction of faulty respiratory sensorial errors.

Thirdly, an increase in faulty detection were noted surrounding periods of frequently missing data. As the M-distance requires all five parameters to be present, any missing parameter will consequently result in a 'missing' M-distance. Following Clifton et al (9) in defining an unstable period to last at least four- in any five-minute frame, frequently missing M-distances could result in this criterium either not being met – or reached too soon - consequently leading to possible faulty conclusions. Partial imputation of missing data could resolve this issue, as often not all five parameters are observed to be missing. For future development, imputing up to two parameters at each moment in time may be considered. Imputation however remains an approximation of unknown data, where imputing more than two parameters may come at the expense of accurate analysis.

Lastly, several stable periods were, wrongfully, detected to be unstable considering patients under heavy active ventilation (*50 breaths/min*). These faulty episodes were considered to be unstable by model A (*combined parameter (in)stability)*, as the RR was deemed too high in relation to other parameters. However, active ventilation can be considered as an iatrogenic fixed value, therefore failing to physiologically adapt when considering other parameters. As stated before, more accurate assessment regarding respiratory capacity at times of active ventilation could come from ventilator-derived parameters. When assessing patients under (heavy) active ventilation, RR may include alternate parameters, or be excluded altogether yet focusing on other vital function variation. Considering patients under ventilation to be sedated, the variation in these vital functions will likely be reduced; the threshold, at which parameters are deemed 'unstable', concerning these parameters may therefore be decreased to account for this reduction in parameter variation.

---

In comparison to the model proposed by Clifton et al(9), several similarities- and differences arise. Although Clifton et al consider an emergency-department setting in adults, their approach to model A *(combined parameter (in)stability)* is similar. Clifton et al consider, among other, the use of a Kernel Density Estimation (KDE) in determining parameters to be (un)stable while avoiding treating the parameters as individual. Where the KDE is resilient to noise, its maximum number of dimensions *(n=4)* prevents its use in our approach of 5 different parameters.

When using the KDE, Clifton et al(9) do treat combinations of parameters at time ***t*** as unrelated to earlier measurements. In our study, we consider each entry as independent *(is my patient stable at this time?),* yet also dependent to earlier measurements *(Is my patient showing vital functions 'normal' to its own curve?)*. Considering a different target population (adults), the dependence to earlier measurements may be limited due to every patient having (roughly) the same vital function values deemed 'normal'. Therefore, we argue our approach to be more specific to the specialized setting of PICU patients with CHD. Most likely, our algorithm applied to an ED-setting in adults may not prove more useful than the proposed approach by Clifton et al.

Considering an ML approach in similar child-intensive care setting, several models have been proposed as reviewed by Hoodhboy et al(14). In this systematic review concerning ML in child- and adolescent health, 6 studies were analyzed where 4 studies were noted to use deep learning of varying complexity. For example, some studies used the stepwise processing of data resemblant to neural processing in the human brain, also known as Neural Networks (NN). These NN are extremely useful in many situations, where their strength lies in robustness, resilience to noise and their capability of handling (very) large datasets. The proposed SVM in model A *(combined parameter (in)stability)* for example, is also resilient to outliers- yet was trained using equal weights for each of its fitted parameters. Arguably, some parameters are allowed to account for more weight than others. For this, a multilayer perceptron (MLP, itself a form of artificial NN) may prove useful as different dimensions are each allowed to carry their own weight. When algorithmic validation is possible in a prospective setup, MLP can be tested against the SVM to analyze whether these different weights will result in an increase in clinical value.

.

## Limitations

There are several limitations to our study.

Firstly, patients were included only when invasive blood pressure and rSo2 was monitored. This introduces selection bias, as not every infant admitted to the PICU with CHD requires an arterial catheter as well as cerebral oxygenation monitoring. Feature expansion of the proposed algorithm, allowing for example both non-invasive blood pressure and invasive blood pressure, may help resolve this bias. Future expansion should focus on features acquired through minimally invasive, frequently applied monitoring. We argue clinical utility likely to increase when less specific situations, i.e., monitoring data, are required to meet the minimal data-requirements in order to make use of the algorithm.

Secondly, the population used for training purposes consisted only of infants under the age of twelve months. With varying parameter values considered 'normal' at different stages in life, especially below the age of twelve, aberration detection would likely result in faulty detection when used in these patients which are not accurately represented in the training dataset. Implementing different age-specific 'base' values, i.e., normalizing parameters towards different values depending on what is deemed 'normal' in CHD patients at the specific age, may further increase clinical value.

Thirdly, only continuous data streams were used in the assessment of patient status. No discontinuous data was added, such as non-invasive blood pressure or laboratory values. These features may provide essential information regarding current patient status, where focus lies in their *added* value to continuous data. We argue the minimal data-requirement to stay as low-key as possible, with reasonable accuracy, to maintain a wide use. Extra (discontinuous) data streams may be beneficial in addition to the 'starter kit' of parameters minimally required to make use of the algorithm.

Fourthly, by using the M-distance in determining the sum of distances for each parameter to their respective means, an assumption is made that any value above- or below their mean is considered to be a negative development. Arguably, considering SpO2 and rSo2, any value above its respective mean need not be a clinically relevant negative development. This issue could partially be resolved by implementing a log-normal-distribution, instead of its gaussian counterpart, upon normalizing specific parameters.

Lastly, as stated in chapter IV, retrospective analysis of patient status was performed by a single expert in the field relying solely on medical notes and vital functions. This approach is not ideal, as the limited information allows room for individual. Ideally, accuracy would be tested prospectively where two members of the medical team are asked to check the patient at times the algorithm is triggered. When in disagreement, a third expert opinion could resolve the issue. This would enable more accurate evaluation and therefore provide additional information regarding clinical use. For example, if triggered; would the medical team have found the alert to be 'late', 'on time' or 'too soon'? As a bonus, the algorithm could be tested in the desired population: PICU patients with CHD *currently* admitted, instead of a retrospective population whose etiology-, surgery- and/or treatment may vary from the patients admitted to date.

## Strengths

Considering a systematic review performed by Hoodbhoy et al regarding machine learning in child-medical settings early 2021 (14), few (*2%, 6 studies)* models described an ICU setting, where their respective models were mainly *(5 out of 6 studies)* designed for either diagnostic purposes or prediction of future events. The 6th study was aimed at infectious control in children admitted to an ICU. Whilst still a concept, proposed algorithm is a first, small, step into using monitoring-aimed machine learning in a PICU setting. As no deep learning was used, where models are at times considered un-customizable '*black-boxes'*, a big strength of this study is its subsequent ability to adjust settings. This flexibility may in the future be aimed towards maximizing clinical usefulness.

Additionally, a lot of effort was put into visualization of algorithm output and analyzing the models at different stages in development. This allowed, as discussed early in this chapter, for several observations which may help improve accuracy in future versions of the algorithm.

## Recommendations

A multitude of studies have started base development of ML aids to detect various events in different settings(13,14). For most of these models, their setup remains retrospective in nature, though ample evidence is presented of its possible clinical usefulness. Recognizing the complex data-infrastructure needed, prospective evaluation of models is arguably the way to move forward. Despite these limitations, some successful efforts have been made in the prospective setup of ML-analysis paving the way for others (9,13).

As stated before, multi-expert evaluation in a prospective setup can be considered as the ideal accuracy evaluation method. At times when the same data is available to each model – resulting in fair comparison – experts may accurately assess which ML-model is optimal for the specific clinical situation (e.g., PICU, ED). We therefore strongly encourage prospective evaluation to enable fair, expert-evaluated comparison between different ML-models, such as MLP, NN and SVM.

Whenever data-driven, evidence based, ML is able to objectively- and accurately detect patient (in)stability, future research may focus on interventions resulting in less periods of instability. In the case of PICU-infants with CHD, this may result in less complications and, ultimately, an improved quality of life due to less morbidity.

Additionally, as stated earlier, we recommend the minimal data-requirements for future models to stay as low-key as possible to maintain a wide use. Future research is encouraged to consider a 'basic' ML setup where additional data streams can be implemented to provide *additional* information, rather than become part of the minimal requirements. This setup could be beneficial for specific patients with intensive, widespread monitoring, yet also allow the algorithm to remain within reach for the more low-key patients requiring minimal monitoring.

Earlier throughout chapter V, several algorithmic adjustments were recommended to increase clinical usefulness which will not be repeated here.

# VI – Conclusion

The described concept of a dual-approach aberration-detection algorithm can be used to automatically classify PICU monitor datasets, though accuracy should be improved and prospectively evaluated. Feature expansion aimed for non-invasive monitoring could further increase clinical value, where ML may eventually prove useful both as an addition to conventional monitoring as well as the research-related processing of big datasets.

# VII – Reference list


1. Sun RR, Liu M, Lu L, Zheng Y, Zhang P. Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments. Cell Biochem Biophys 2015 723 [Internet]. 2015 Feb 1 [cited 2022 Jan 11];72(3):857–60. Available from: https://link-springer-com.proxy.library.uu.nl/article/10.1007/s12013-015-0551-6

2. Van Der Bom T, Zomer AC, Zwinderman AH, Meijboom FJ, Bouma BJ, Mulder BJM. The changing epidemiology of congenital heart disease [Internet]. Vol. 8, Nature Reviews Cardiology. Nature Publishing Group; 2011 [cited 2022 Jan 11]. p. 50–60. Available from: https://go-gale-com.proxy.library.uu.nl/ps/i.do?p=AONE&sw=w&issn=17595002&v=2.1&it=r&id=GALE%7CA245541738&sid=googleScholar&linkaccess=fulltext

3. Fister P, Robek D, Paro-Panjan D, Mazić U, Lenasi H. Decreased tissue oxygenation in newborns with congenital heart defects: a case-control study. Croat Med J [Internet]. 2018 Apr 1 [cited 2022 Jan 11];59(2):71. Available from: /pmc/articles/PMC5941290/

4. Davidson J, Schaffer MS. Cyanotic heart disease. In: Berman's Pediatric Decision Making [Internet]. StatPearls Publishing; 2011 [cited 2022 Apr 2]. p. 537–41. Available from: https://www.ncbi.nlm.nih.gov/books/NBK500001/

5. Mandalenakis Z, Giang KW, Eriksson P, Liden H, Synnergren M, Wåhlander H, et al. Survival in Children With Congenital Heart Disease: Have We Reached a Peak at 97%? J Am Heart Assoc [Internet]. 2020 Nov 14 [cited 2022 Apr 3];9(22). Available from: https://pubmed.ncbi.nlm.nih.gov/33153356/

6. Barkhuizen M, Abella R, Vles JSH, Zimmermann LJI, Gazzolo D, Gavilanes AWD. Antenatal and Perioperative Mechanisms of Global Neurological Injury in Congenital Heart Disease. Pediatr Cardiol [Internet]. 2021 Jan 1 [cited 2022 Apr 3];42(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33373013/

7. Kumar N, Akangire G, Sullivan B, Fairchild K, Sampath V. Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront. Pediatr Res [Internet]. 2020 Jan 1 [cited 2022 Feb 8];87(2):210. Available from: /pmc/articles/PMC6962536/

8. Barfod C, Lauritzen MMP, Danker JK, Sölétormos G, Forberg JL, Berlac PA, et al. Abnormal vital signs are strong predictors for intensive care unit admission and in-hospital mortality in adults triaged in the emergency department - a prospective cohort study. Scand J Trauma Resusc Emerg Med. 2012 Apr 10;20.

9. Clifton DA, Wong D, Clifton L, Wilson S, Way R, Pullinger R, et al. A large-scale clinical validation of an integrated monitoring system in the Emergency Department. IEEE J Biomed Heal Informatics. 2013;17(4):835–42.

10. Chapman SM, Maconochie IK. Early warning scores in paediatrics: an overview. Arch Dis Child [Internet]. 2019 [cited 2022 Apr 3];104(4). Available from: https://pubmed.ncbi.nlm.nih.gov/30413488/

11. Kowalski RL, Lee L, Spaeder MC, Randall Moorman J, Keim-Malpass J. Accuracy and Monitoring of Pediatric Early Warning Score (PEWS) Scores Prior to Emergent Pediatric Intensive Care Unit (ICU) Transfer: Retrospective Analysis. JMIR Pediatr Parent [Internet]. 2021 Jan 1 [cited 2022 Apr 3];4(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33547772/

12. Johnson AEW, Ghassemi MM, Nemati S, Niehaus KE, Clifton D, Clifford GD. Machine Learning and Decision Support in Critical Care. Proc IEEE Inst Electr Electron Eng [Internet]. 2016 Feb 1

[cited 2022 Mar 29];104(2):444. Available from: /pmc/articles/PMC5066876/

13. Muralitharan S, Nelson W, Di S, McGillion M, Devereaux PJ, Barr NG, et al. Machine Learning–Based Early Warning Systems for Clinical Deterioration: Systematic Scoping Review. J Med Internet Res [Internet]. 2021 Feb 1 [cited 2022 Mar 29];23(2). Available from: /pmc/articles/PMC7892287/

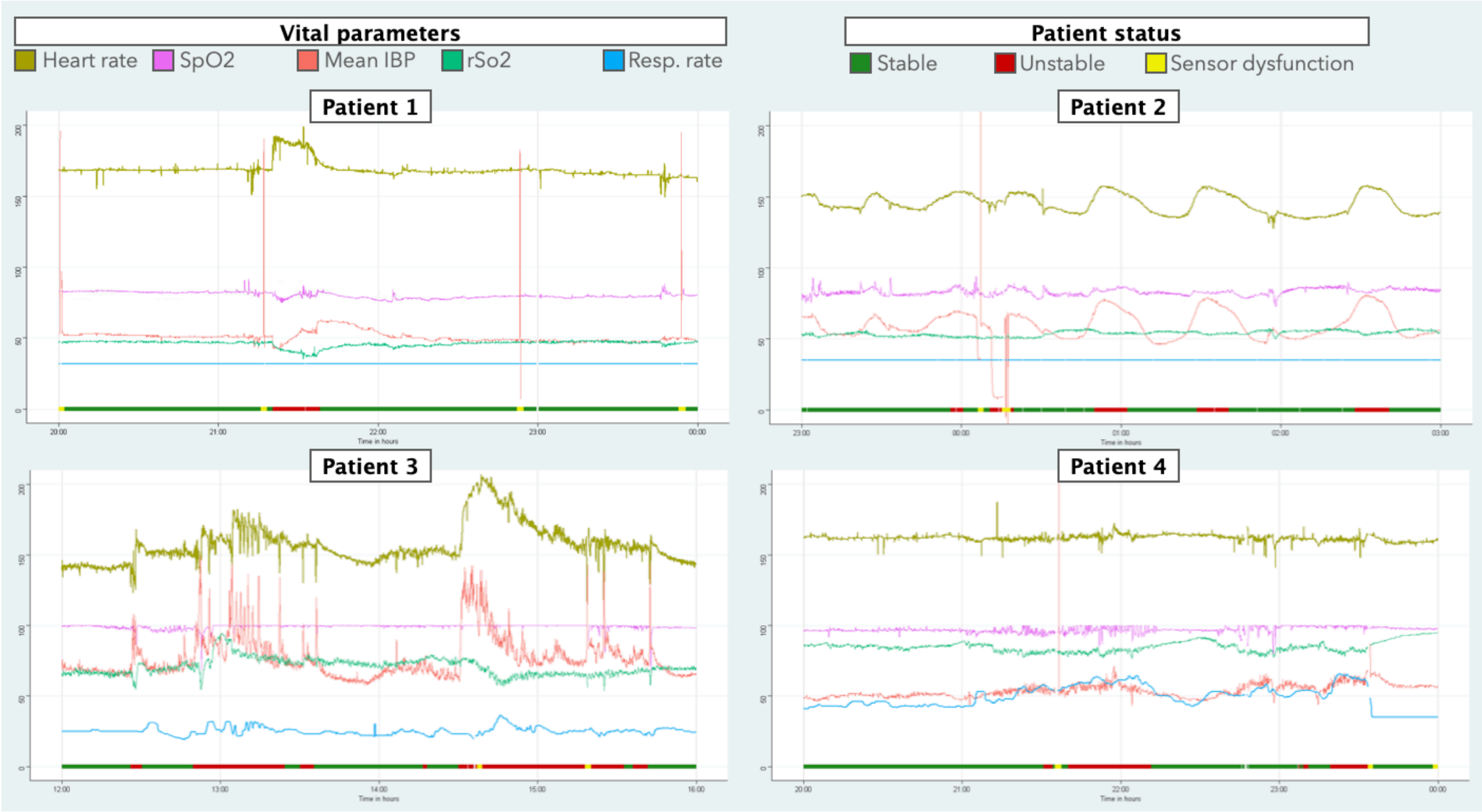
14. Hoodbhoy Z, Jeelani SM, Aziz A, Habib MI, Iqbal B, Akmal W, et al. Machine learning for child and adolescent health: A systematic review. Pediatrics [Internet]. 2021 Jan 1 [cited 2022 Apr 3];147(1):2020011833. Available from: /pediatrics/article/147/1/e2020011833/33441/Machine-Learning-for-Child-and-Adolescent-Health-A

15. De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemom Intell Lab Syst. 2000 Jan 4;50(1):1–18.

# VIII – Appendix

A) Aberration detection visualization in four test patients.
B) Inter-parameter correlation of subgroups Alpha- & Beta.

# Appendix A: Novelty detection visualization in four test-patients



# Appendix B: Inter-parameter correlation.

**Pearson correlation on complete data, similar results using Spearman's correlation*

| *n=52* | **Sat > 90%** | Heart rate | Resp. Rate | Mean BP (invasive) | rSo2 | SpO2 |
|---|---|---|---|---|---|---|
| | Heart rate | - | 0,12 | -0,09 | -0,21 | -0,06 |
| | Resp. Rate | 0,12 | - | *0,00* | -0,02 | *0,01* |
| | Mean BP (invasive) | -0,09 | *0,00* | - | 0,17 | 0,08 |
| | rSo2 | -0,21 | -0,02 | 0,17 | - | 0,17 |
| | SpO2 | -0,06 | *0,01* | 0,08 | 0,17 | - |

| *n=26* | **Sat < 90%** | Heart rate | Resp. Rate | Mean BP (invasive) | rSo2 | SpO2 |
|---|---|---|---|---|---|---|
| | Heart rate | - | 0,04 | -0,06 | -0,11 | -0,20 |
| | Resp. Rate | 0,04 | - | 0,05 | -0,01 | -0,10 |
| | Mean BP (invasive) | -0,06 | 0,05 | - | -0,08 | 0,14 |
| | rSo2 | -0,11 | -0,01 | -0,08 | - | 0,40 |
| | SpO2 | -0,20 | -0,10 | 0,14 | 0,40 | - |