MIG FORENSIC - REPORT

S.DUBAIL, P.PILI , A.AGREBI, J.AMAR, P.BERRANGER, E.COTET, T.DUBREUIL, A.GILLIET, J.JAHAN DE LESTANG, A.LAFORGUE, A.LEUBA, E.MARES, , K.PROVOST, T.RENAUDIE, M.STOLL, R.ZAGHROUN

SUPERVISORS : Sébastien TRAVADEL, F.GUARNIERI, D.DELAITRE, Xavier ALACOQUE, Mathis BOURDIN

I. INTRODUCTION **3**

I.1. The importance of data in the hospital? 3
I.2. What does usual pediatric surgery look like? 3
I.3. The MIG-FORENSIC approach 6

II. METHOD **7**

II.1. Observation methods in the operating room 7
II.2. Data Processing 7
II.2.a Pre-processing *8*
II.2.b. The labelling issue *11*
II.2.c. Feature engineering and extraction *12*
II.2.d. Analysis Algorithms *20*

III. RESULTS AND LIMITS **26**

III.1. Understanding of the results 26
III.2. Algorithm and decision making 27

IV. CONCLUSION **29**

VI. APPENDIX **30**

Appendix.1 List of the surgeries we attended: 30
Appendix 2 Scheme of the medical reaction according to what the algorithm responds 31

VII. BIBLIOGRAPHY **33**

VIII. SPECIAL THANKS **34**

I. Introduction

I.1. The importance of data in the hospital?

In the hospital, very little data is collected, and most of the time it is impossible to analyze it properly due to the lack of unity between the different sources. During the past decade, efforts have been made to collect data at a higher scale and to centralize it, but there is still a very long way to go.

However, data analysis is ubiquitous in medical research. During a visit to the research center in oncology of Toulouse, several research teams, including a biologist, a doctor and a data scientist, presented us the way they use machine learning algorithms in their work. They used high quality data which allowed them to carry out high level research studies.

In the operating room, data from the monitor is usually only available instantaneously and is not saved. Instead, the evolution of the parameters is reported on an anesthesia sheet that is kept. Information technology can sometimes be overlooked in the medical field; thus, it is relevant to wonder if data has a role to play in the operating room.

While statistical learning is now an integral part of the engineering landscape, the approach of delegating all or part of the decision to an algorithm is still not widespread in surgery. This can be explained on the one hand by the limited amount of data available; and on the other hand, by the difficulty to develop a "responsible AI" in the broad sense, i.e. to answer the ethical questions raised by unreliable algorithms used for critical decisions. The case under study, submitted by the pediatrician surgery of the university hospital center of Toulouse, addresses those two aspects. It consists of both : to first develop classification algorithms to predict cardiac arrests during pediatric surgeries; and studying how these tools may be accepted in a professional setting.

I.2. What does usual pediatric surgery look like?

In the surgical service of the pediatric hospital, there are 9 operation rooms. 5 out of 6 are usually used simultaneously. Each one has a specialty among which, cardiac surgery, orthopedics, visceral surgery, etc. The regulators make the schedule for all the doctors, juniors, and nurses.

Before the patient's arrival, the anesthetic team (usually made up either by one junior doctor or a nurse and a supervising anesthesiologist) sets up the tools for anesthesia. They agree on a protocol for the patient and set up the monitoring devices.

The patient arrives and is transferred on the operating table. The anesthetic team sets up electrodes on their torso and an oxygen saturation (SpO_2) sensor (on one finger. Young children are put down to sleep by the anesthetic team using a special gas. The SpO_2 indicator refers to the amount of oxygenated hemoglobin in the blood. Only then, an intra veinous (IV) is set up to inject Propofol to keep the patient asleep. Older children are directly being injected Propofol, and the IV is set up before sleeping. Analgesics are always used, while paralyzing agents like curare are not always needed. When the patient is asleep, more

invasive sensors can be set up, like in the femoral artery. The patient can also be intubated; it is usually the case for newborns.

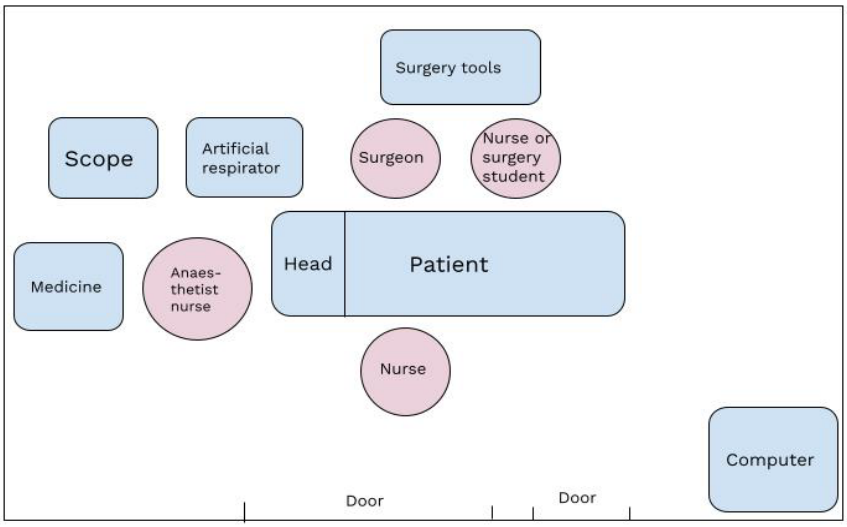
When the anesthetic team is done with the patient, the surgeon is called. The information about the patient is checked. The surgeon incises the operating site and goes on with surgical gestures. They are helped by the junior surgeon and a nurse. The three are scrubbed in as opposed to the rest of the people in the operation room (OR). During the entire surgery, the patient is monitored by a member of the anesthetic team – a junior doctor or a nurse if the surgery is basic, the attending physician if the surgery is risky. When the surgery is basic, the attending physician navigates between about three operating rooms. Monitoring devices give information like cardiac rhythm, blood pressure, body temperature, SpO_2 and pCO_2. The anesthetic team sets up an alarm threshold for each of those parameters. When an alarm goes off, the anesthetic team silences it and analyzes the situation. Alarms are frequent and most of the time unsignificant.

When the surgery is over, they leave the operation room, and the anesthetic team takes the lead. The anesthetist and its junior will agree on a protocol to wake the patient up and manage post-surgery care. Critical patients are directly transferred to the Pediatric Intensive Care Unit (PICU) or the Neonatal Intensive Care Unit (NICU). Otherwise, general anesthesia is interrupted, but analgesics are maintained. The anesthetic team looks after the patient during awakening and stimulates them during the last stage. As soon as the patient shows signs of consciousness, they are brought to the recovery room.

If another surgery is scheduled right after, the rest of the anesthetic team prepares the operation room for the next patient. Special caregivers thoroughly clean the operating room while the scrub nurse removes all the sterilized elements previously used.

figure I.1 An operating room during a surgery

figure I.2 Operating room structure



The anesthetist and the surgeon hardly ever communicate apart from the beginning and the end of the operation. Indeed, those two periods are delicate as the patient is either put to sleep or awakening.

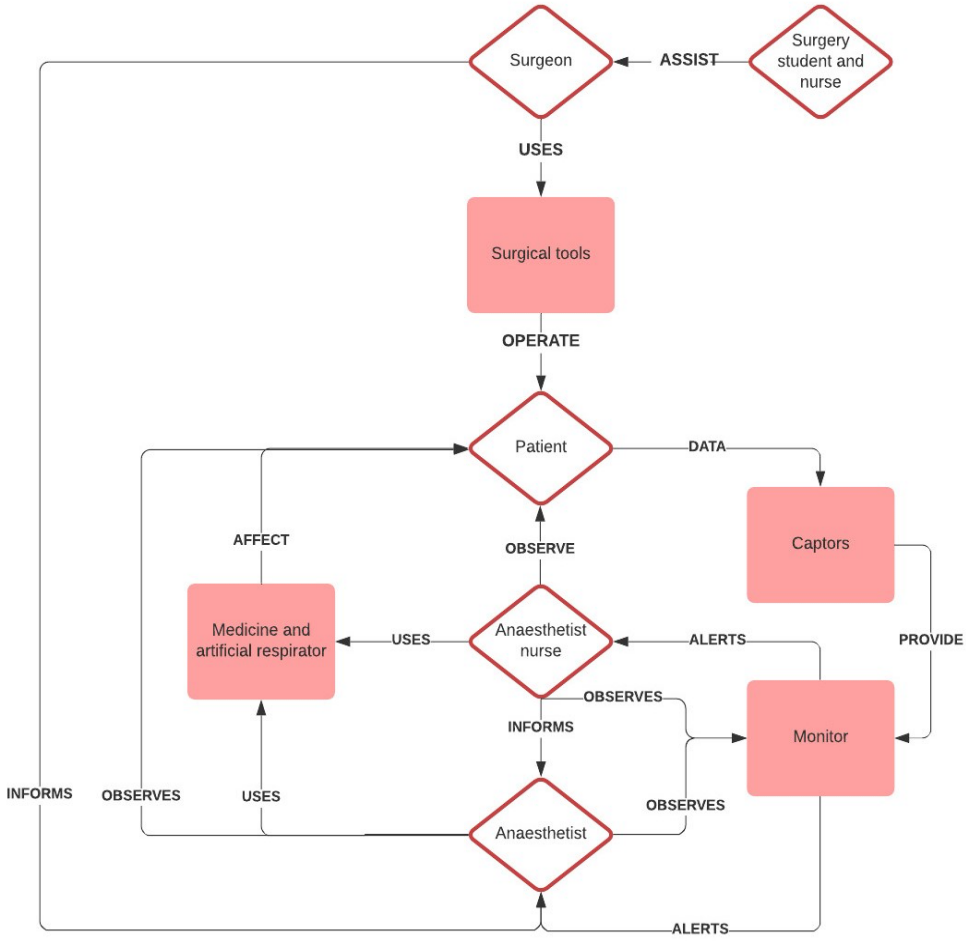


figure I.3 Operating room interaction

I.3. The MIG-FORENSIC approach

The MIG FORENSIC studies decision support in real-time surgery according to two components, explored in parallel to submit an overall vision of the problem. The first one is the development of Machine Learning algorithms based on real data. The second aspect is both the study of the implementation of the solution in operating rooms and in the decisional process, as well as the analysis of data. We will also tackle ethical issues raised by the introduction of such tools, by confronting the performance of the algorithms with the importance of the decision. To study these two objectives, the group has been split up into two parts :

The FORENSIC-DATA Group: This group will study actual patient monitoring data, recorded during several months in six operating rooms. This corpus of data has unique characteristics that make it difficult to analyze: limited number of cases, multiple pathologies, varied population (from babies to young adults). It is about taking control of this data through the entire processing chain, in order to develop the first classification algorithms capable of predicting heart failure during major cardiac insufficiencies likely to seriously disrupt the course of the surgery or even threaten the life of the patient. Signal processing, statistical learning and software engineering courses during the first week have allowed a progressive approach to this field.

The FORENSIC-Ethics Group: The other group has conducted non-participatory observations in operating theatres (Toulouse University Hospital, pediatric surgery center). The objective was to model the decision process during the surgical act, and to determine under which conditions and to what extent this decision could be assisted by a tool based on statistical learning. The experts' confidence in such a tool especially depends on its long-term reliability, and its intelligibility when the prediction contradict intuition. However, most of the time, the performance of classification algorithms increases at the cost of a significant complexity, until they turn into "black boxes". This is particularly true of Artificial Intelligence algorithms, which have achieved the performance of the best experts. The students thus had to answer the following question: how can we consider implementing a solution into operating rooms that would not violate the experience of the doctor?

II. Method

II.1. Observation methods in the operating room

In the operation room, we were really focused on how the whole system worked, to understand what the role of each actor was. The idea was also to discover the relative importance of each monitoring parameter and to learn about some technical problems that the algorithm will face. For instance, arterial pressure monitoring is disturbing $Sp0_2$ monitoring, sensors are sometimes poorly placed, the electric lancet is disturbing cardiac frequency monitoring, etc… We have also learnt a lot about monitoring aberrations that can be useful to understand data properly.

Our goal was to map the decisional process in the operating room. The idea was to find how our solution could be implemented, and at the same time accepted. We wanted to find out where the solution might stand in the decisional process. To do so, we were highly focused on each micro-event, writing down the time it happened and what the anesthetist did. Then we asked the anesthetist what happened and why he reacted the way he did.

We wanted to determine the perfect form for the tool: which display, what it should tell… The idea was to materialize the solution in order to be implemented. The method we used was to wait until the drowsiness of the patient to ask questions that we had prepared upstream. There are a few questions that we kept asking to every anesthetist we met. For example: what is according to you a good tool? How long in advance the algorithm should tell you about an anomaly in order to be useful? Should it include sound alarms? The answer of the following questions will be discussed with respect to the results of our algorithm.

II.2. Data Processing

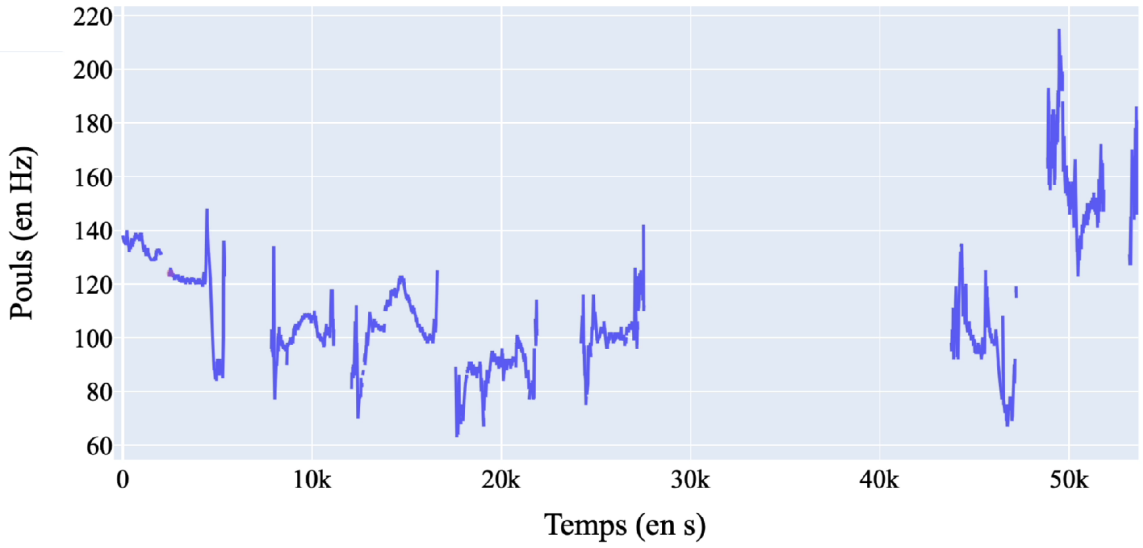
Machine Learning is the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. One of its most widely used tools is data analytics and in particular predictive analytics. Predictive analytics uses a single tool, the classifier, to solve a wide range of problems. One of these problems is to determine, from its characteristics, the state of a hidden situation by automatically classifying the data into one or more sets of classes.[2]

We worked on all the following steps, first on the preprocessing of the data. During this phase, the raw data are carefully checked for possible errors. The objective is to eliminate poor quality data, i.e. incomplete or incorrect data. The data must be made usable for further processing. In a second step, we must label the data, which corresponds to assigning a class to each data. Then comes the Feature Engineering. This is a process that consists in transforming the raw data into features that more accurately represent the problem underlying the predictive model. In simple terms, it involves applying domain knowledge to extract analytical representations from the raw data and preparing them for Machine Learning. Only those features are then given to the classifier. It is then necessary to test different classifiers to find the best one. This quality varies from one problem to another.

II.2.a Pre-processing

First we received months of raw data from the CHU of Toulouse's operating rooms and we had to preprocess our data to clean it, and remove useless indicators.

Figure II.1 Cardiac frequency of a raw data sample



II.2.a.i. Mapping

As we have seen, patient's data are incomplete, and full of gaps. And, even if it were complete, we couldn't keep hundreds of characteristics. This would be too much for the classifiers. Thus, we needed to select the characteristics that were the most important and that were also quite complete for most of the patients. Otherwise, even the gap filling process wouldn't have produced coherent results. This is the role of the mapping function.

By looking at the data, trying to figure which characteristics were recorded more often than the others, and by talking with an anesthetist, we eventually decided which characteristics we were going to keep. Indeed, we kept only the pulse, the oxygen saturation, the pressure, the temperature, and the respiratory frequency of the patient.

Moreover, in the data that the hospital was sending, two columns correspond to the pulse: one measured directly on the patient, and the other deducted it from the oxygen saturation. Those two characteristics were relatively close to each other, so we decided to merge them, by copying the first one and, when it wasn't defined, completing it with the second one. Thus, we created only one pulse characteristic, which was more complete than the original. We did the same thing for the pressure and temperature. After the mapping function is applied, the patient data is composed of only five characteristics that are as complete as possible. In fact, when we built the features, we decided to manipulate only the pulse and the oxygen saturation, hence, three of those five characteristics have not been used whatsoever, above all because they were still not recorded for every patient.

II.2.a.ii. The splitting process

As each operation room file could contain several surgeries, we had to split the raw files into smaller patient files in order to classify them.

The first step of the splitting process was to understand the data we were given. In fact, it appeared that the duration time of each operation could vary. Consequently, we chose to cut operation room files based on cleaning duration between surgeries, to reduce drastically the variability we presented before. Indeed, this parameter is more standardized in the surgical unit: the threshold used was 20 minutes of absence of data registration(in red in figure II.2).

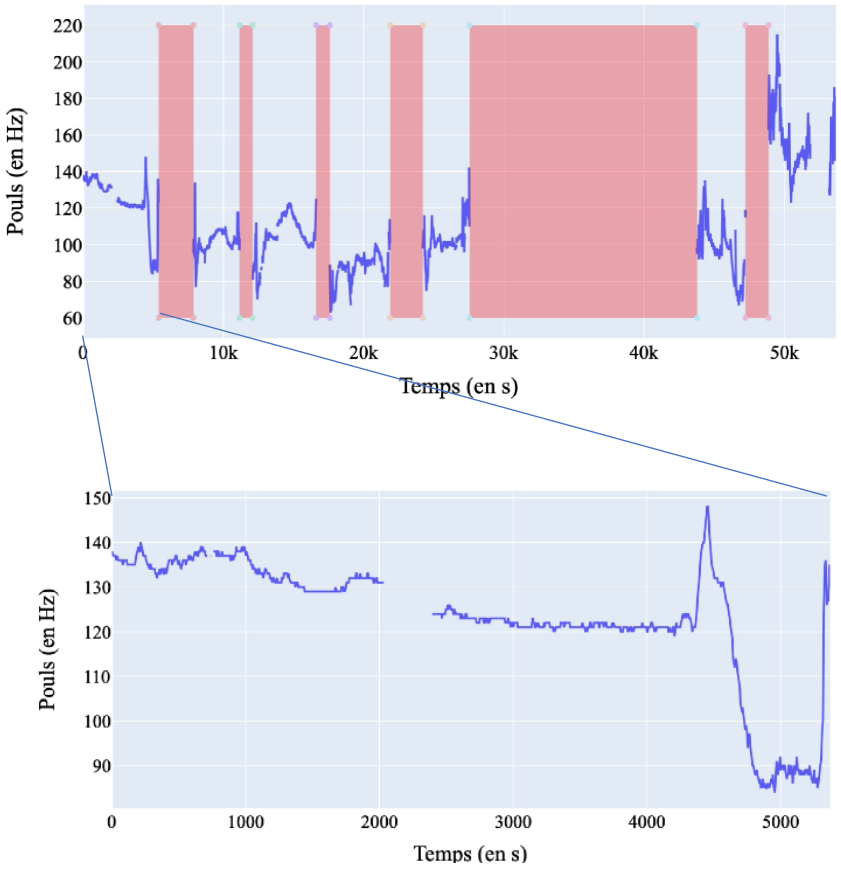


Figure II.2 From a raw data sample to a single patient

The next step towards useable data was to complete missing data through gap filling methods, and to smooth it for further analysis purposes. When an anomaly or an attack is to be observed, we cut it out of the patient data so that the prediction algorithm can learn on data which doesn't include the dangerous event itself (in red in figure II.4), but the previous behavior which may have led to it.

In order to use the features, and especially the ones with the Fourier transform, the functions that represent the evolution of the pulse as a function of time must be differentiable. However, the data given to us comes with three problems :

First of all, the captors give sometimes wrong measures, which leads to what is called “outliers”. These are points that break the regularity of the curve during a very short time and must not be lumped together with sudden variations of the curve that are not mistakes and that contain information. Such outliers create new frequencies in the Fourier transform that should not be here and therefore must be deleted. In order to do this, we used the functions rolling mean and rolling standard deviation. These functions take a panda series as an argument and calculate respectively the mean and the standard deviation around each point using a certain window (we chose 20 points around the point where the mean and the standard deviation are calculated).

Another problem is that the sensors do not deliver any measure sometimes, which creates some gaps in the curve. To tackle this issue, we use interpolating functions which do not modify the curve where there is no gap, and just fill the gaps by using a certain type of interpolation : we chose linear interpolation.

The last problem is the angular points on the curve. Since there is only one measure each 5 seconds, a short variation creates angular points, which need to be tackled otherwise the curve will not be differentiable. To deal with this issue we used a convolution with a “hamming window”, which means we consider the result of a convolution between the curve and a Gaussian function. It is a famous way to smooth any curve.

The result of these three methods is a new curve that has the same values as the first one but without the outliers, with the gaps filled and which has smoother variations between the different measures in order to eliminate the angular points.

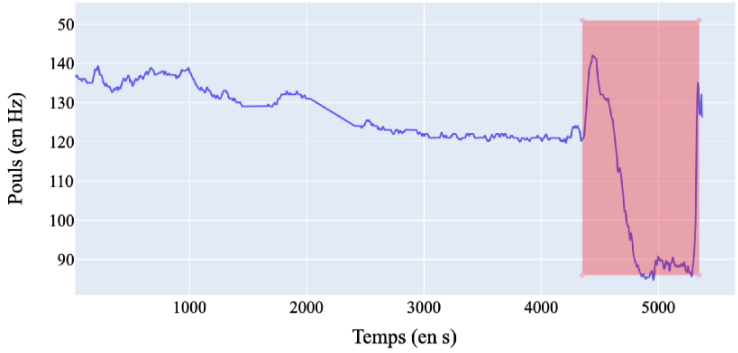


Figure II.3 Single patient smoothed data containing an attack

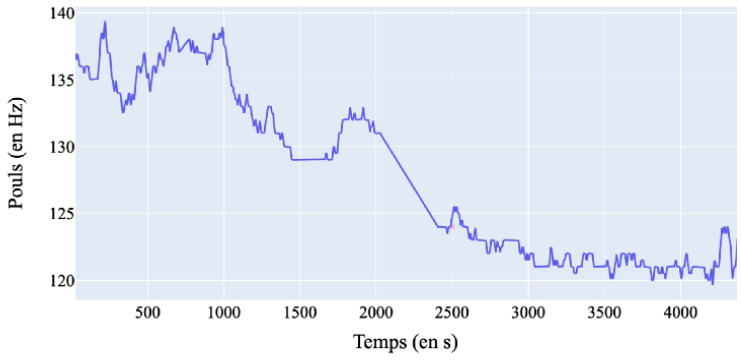


Figure II.4 Final single patient data

II.2.b. The labelling issue

One of the problems we faced during this project was the lack of precise information about the outcome of the surgery, or whether it went well according to the medical staff. To train our algorithms, we had to manually classify each patient in different categories: attack, anomaly and clean. While the 'clean' category means the patient is safe, both 'anomaly' and 'attack' categories are associated with a certain level of risk.

To classify, we must understand how a surgery works. First the patient is put asleep by the anesthetist, this corresponds to the induction period. At the end of the surgery, the patient is woken up. Those two periods correspond to brutal variations of heart frequency we hence must classify as normal. Then, we followed Doctor Alacoques's advice to label the data. We looked at the way the heart behaves through the variations of frequency and $Sp0_2$. Here are a few examples:

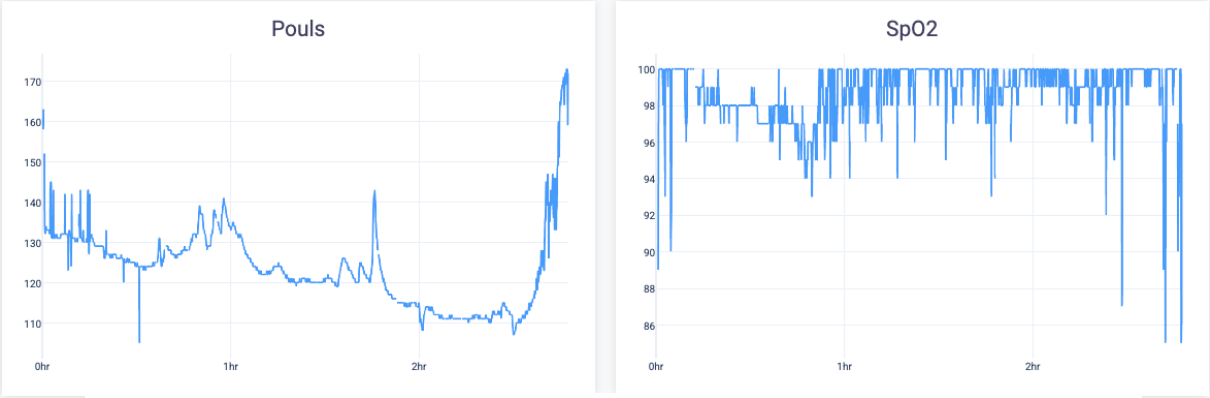


Figure II.5 clean patient data

We decided to put this patient in the clean category because apart from the critical periods, his SpO_2 and heart frequency remain stable during the surgery.

Figure II.5 attack patient data

We classified the previous curve as an attack because there is both a drop in heart frequency and in SpO_2, which is the symptom of an attack. We are particularly careful when covariations are observed.

Figure II.5 anomaly patient data

This patient is classified as an anomaly because even though a drop of SpO_2 is clear, it does not correspond to a significant variation of frequency.

Such a classifying method lacks precision and objectiveness; indeed, labels were often debated for each patient.

II.2.c. Feature engineering and extraction

In order to enable a predictor to classify a patient's data into one of the categories we defined, we need to provide it with quantified characteristics which represent the given data.

The most simple ones are to be extracted from the vitals themselves: statistics such as the mean, standard deviation, skewness, kurtosis, quartiles, minimum and maximum values.

However, restricting ourselves only to these features would make us miss a lot of crucial information, hidden deeper in the data, and which need further analysis to be extracted. We therefore turned ourselves to time-frequency signal analysis.

II.2.c.i.. Features extracted from Fourier Analysis

Continuous Fourier Transform[11]:

A Fourier transform is a mathematical transform that decomposes functions depending on space or time into functions depending on spatial or temporal frequency.

The Fourier transform of a function of time is a complex-valued function of frequency, whose magnitude (absolute value) represents the amount of that frequency present in the original function, and whose argument is the phase offset of the basic sinusoid in that frequency.

$$\mathrm{F}(\nu) = \mathrm{TF}[f](\nu) = \int_{-\infty}^{+\infty} f(t) e^{-2i\pi\nu t} dt$$

Discrete Fourier Transform (DFT):

The vitals used for the analysis being represented as a finite set of uniformly spaced time-samples, the Fourier Transform needs to be discretized in order to compute the frequency spectrum. DTF therefore transforms a sequence of N complex numbers

$$\{x_n\} = \{x_0, x_1, .., x_{N-1}\}$$

into another sequence of complex numbers

$$\{X_k\} = \{X_0, X_1, .., X_{N-1}\}$$

defined by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{i2\pi}{N} kn}$$

Short Time Fourier Transform:

Although being a massively used transform, DFT has its flaws. This method of signal analysis takes as an input the entire sequence of data, and therefore cannot provide any temporarily located information. So is the idea of a Short-Time Fourier Transform (STFT). It is used to determine the sinusoidal frequency and phase content of a signal while it is changing over time, and thus obtain the variations of the Fourier Transform analysis.

The idea is to switch from the time domain to the frequency one to detect behaviors which were not accessible on the temporal signal, such as eigenfrequencies or spectrum energies. It therefore allows us to quantify the excitation of the signal on a frequency level, at a given time, as an analysis window is sliding along the signal and computes such features as discretized function of time.

General hypothesis used to compute the Fourier features:

As we compute features from vitals which are not all of the same size, we need to define the same time window to apply the STFTs for all signals. As samples length can stretch from thirty minutes to several hours, with an induction time of fifteen minutes, two windows seemed appropriate: from the beginning to 35 minutes of surgery (therefore including induction) and from the end of the induction to 35 minutes of surgery. The induction is indeed a period of time where the vitals are particularly agitated, the idea was therefore to let the door open for us to add it or not to the data the predictors will work on.

In practice, the procedure for computing STFTs is to divide a longer time signal, here defined by the time window chosen (largest orange window in figure II.6-7), into

shorter segments of equal length and then compute the Fourier transform separately on each shorter segment. This reveals the Fourier spectrum on each shorter segment.

To define the length of such segments (smallest orange window in figure II.6-7), a compromise must be made between temporal and frequency definition. Indeed, a too short time-segment implies less data for the Fourier transform to be computed on, and thus a frequency analysis less precise, whereas a too long time-segment would imply a less local analysis of the signal. The frequencies axis has also been limited to 0,04 Hz, as the coefficients for further frequencies were unsignificant.

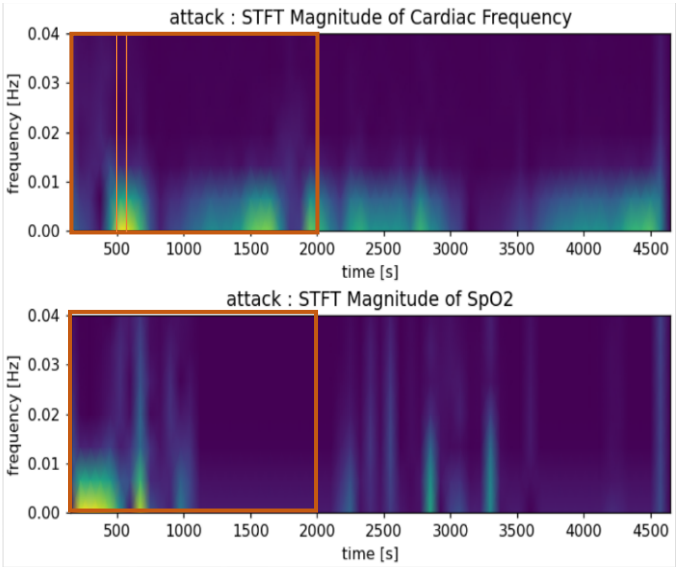


Figure II.6 STFT analysis for a segment length of 150s

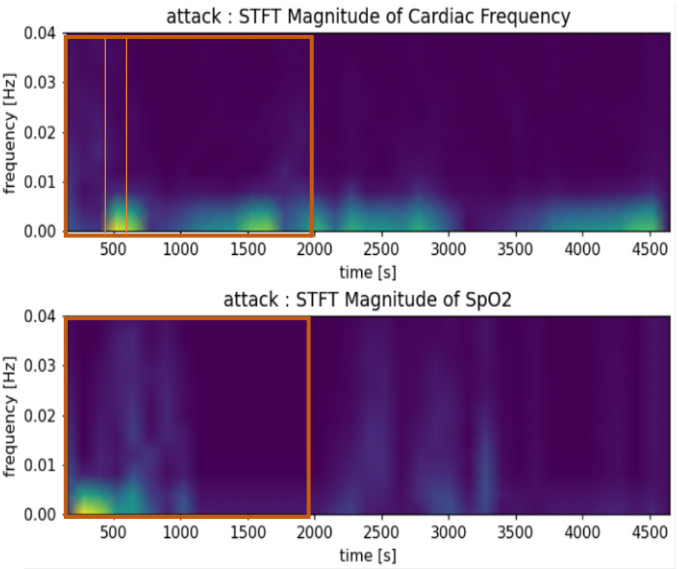


Figure II.7 STFT analysis for a segment length of 250s

Here are two examples for 150 seconds and 250 seconds length segments : given the global time window the Fourier Transform will be computed on, a good compromise appears to be 250 seconds length segments.

First feature group: spectrum energies and their correlation:

What can be first observed is the variation of the coefficient's magnitude on the STFT : for some data, the higher they are for the cardiac frequency, the lower they are for the oxygen saturation : therefore some correlation was computed. In order to quantify such observations, the coefficients are integrated along the frequencies axis to obtain the time variation of spectrum energies for both cardiac frequency and oxygen saturation. We then proceeded to compute a Pearson correlation coefficient between the two.

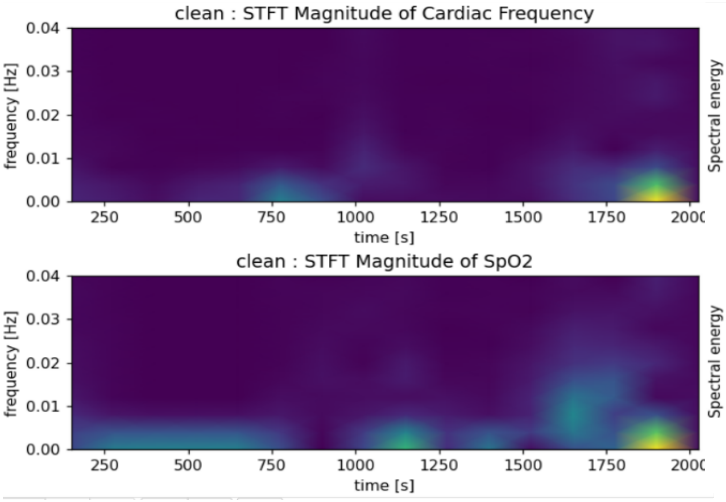


Figure II.8 STFT analysis.

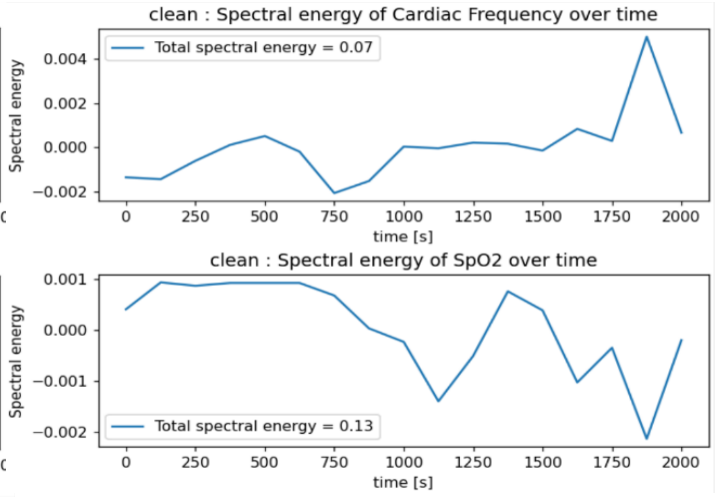


Figure II.9 Corresponding spectrum densities.

To quantify globally the signal's "excitation" during the chosen time window, we also integrated the spectrum energies over time to obtain a second feature, indexed as « Total spectral energy » on figure II.9.

However, by integrating so, we lose crucial information about the frequencies of the spectrums. We therefore thought of a way to quantify a link between frequencies of the cardiac frequency and the SpO2, as it could represent relevant information about the patient's state.

Second feature group : SpO2/FC correlations for given frequencies:

Given the fact that STFT discretizes the limited frequency domain into 26 segments, we computed for each segment the correlation between the associated Fourier coefficient magnitude of the cardiac frequency and the SpO_2. The statistical overview of the results enabled us to choose three frequencies which stood out more than others, regarding their capacity to statistically classify a patient: their correlation coefficient values differ between anomaly patients and attack/clean ones,

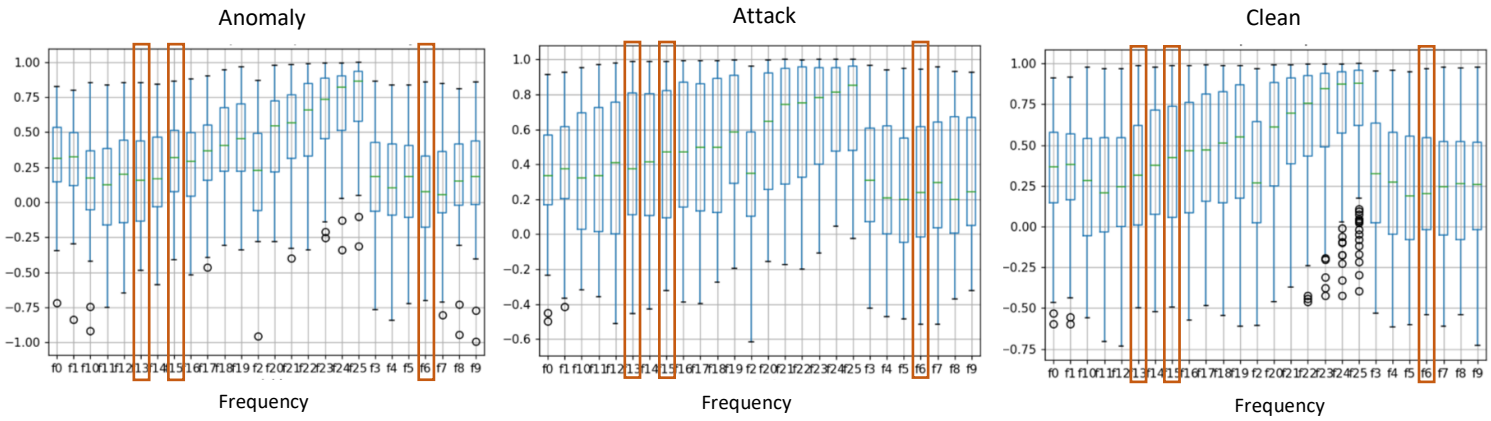


Figure II.10 Statistical overview of the correlation coefficient by frequency, for each

II.2.c.ii.. Features extracted from wavelets analysis

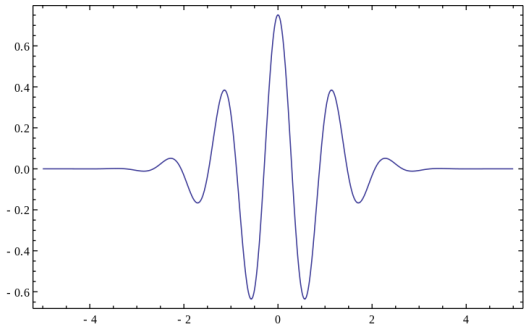
Another essential feature engineering route rested on wavelet analysis. Wavelets are brief, wave-like oscillations that begin at zero, oscillate, then return to zero. Wavelets can be used to extract information from a signal.

Although FFT is only localized in frequency, wavelet transform is localized in both time and frequency.

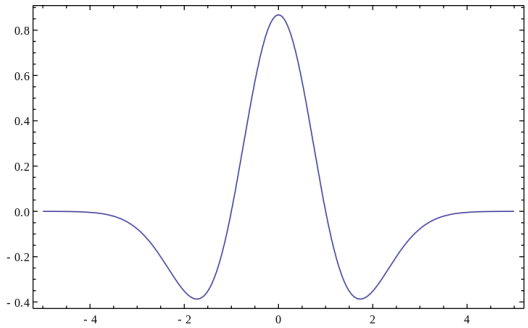
Continuous wavelet transforms (CWT)[9]:

Continuous wavelet transforms are the projection of our signal onto a continuous function family of frequency bands [f, 2f]. The first frequency band is given by the "mother" wavelet $\psi(t)$. For example.

Morlet Wavelet



Mexican Hat Wavelet

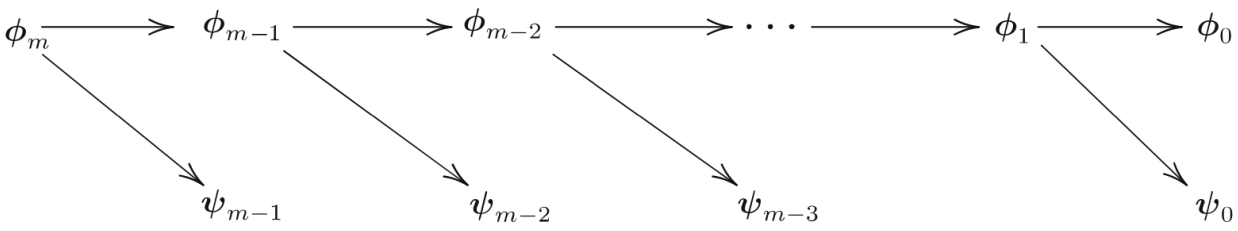


The next frequency bands are scaled versions of this wavelet. Frequency band $[1/a, 2/a]$ is generated by the "child" wavelet: $\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$

The projection of our signal onto these sub-bands is: $WT_{\psi}(a,b) = \langle x, \psi_{a,b}\rangle = \int_R x(t)\psi_{a,b}(t)dt$

Where the wavelet coefficients are: $x_a(t) = \int_R WT_{\psi}(a,b)\cdot\psi_{a,b}(t)db$

The m-level DWT used is defined as the change of coordinated from ϕ_m the original function to $(\phi_0, \psi_0, \psi_1, \ldots, \psi_{m-1})$, where ϕ_0 is the "approximation" of low-est frequency to ϕ_m, and ψ_0 to $\psi_{m-1}$are the "detail" (higher frequency) functions, with ψ_{m-1} being the highest frequency function. In practice, we can resort to a discrete wavelet transform (DWT), using a discrete subset of frequencies.



Thus, both the heart rate (HR) and the oxygen saturation of the patient was decomposed onto these various frequency sub-bands. This allowed for comparisons of the two signals.

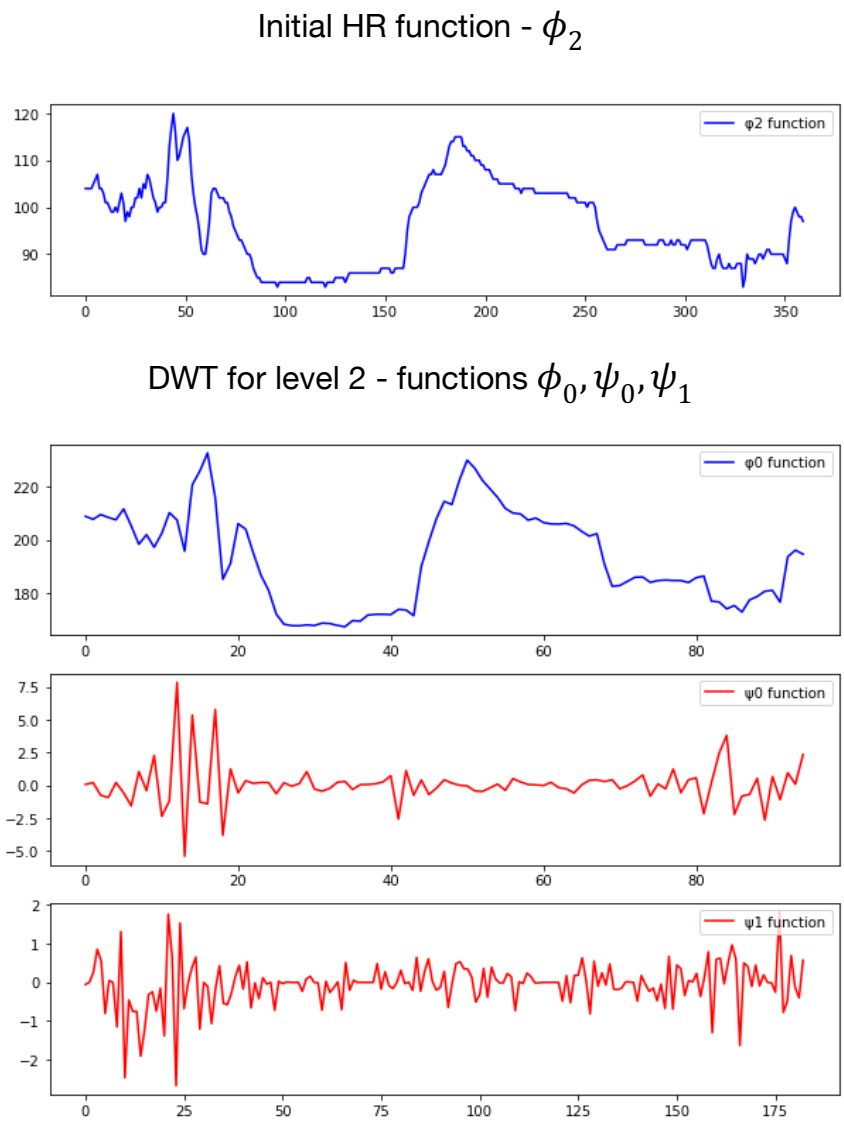


Figure II.10 Successive levels of decomposition .

We supposed there to a distinctive correlation between SpO2 and HR. Namely, that SpO_2 variations could give birth to HR variations. Since higher frequencies are indicative of the reactions of a given signal, we chose to analyze the correlations between the lower frequencies of SpO_2 (notably its ϕ_0 function) and the higher frequencies of HR (such as its ψ_0, ψ_1 functions, in red above). We observed, as a whole, that "clean" patients seemed to react more strongly to the ups and downs of their oxygen intake.

Another computed feature was the signal energy density for several levels of wavelet decomposition.

Wavelet features were therefore invaluable for the classifiers.

II.2.c.iii. Principal Component Analysis

The PCA is a machine learning method that processes a signal with autocorrelated points to determine the mutually independent components that best summarize the information.

The goal is to reduct the dimension (ie. number of time points) of a patient's surgery and check if we can reliably understand the variation of the cardiac frequency with only a few points.

In practice, we trained the PCA with a normal set of clean patients on the first twenty minutes of a surgery (our initial dimension is n=200 (200 x 5 sec = 20 minutes)). The algorithm uses the covariance matrix of the training set , the algorithm finds $k < n$ new principal components which are the components of maximal variance of the set whose information is greater than a threshold that we fixed at 90% of the initial information. Thus, the training set provides a new matrix of projection from the n-dimension space to the new k-dimension space. We found that for k=25 (approximately two minutes of surgery) we could resume 95% of the next twenty minutes.

When we took a new measure on which the algorithm was not trained, the measure is transformed by being projected in the k-dimension space and then projected back in the n-dimension space.

This approach is both geometrical (variables are represented in a new space) and statistical (the research of the best points to maximise the variance).

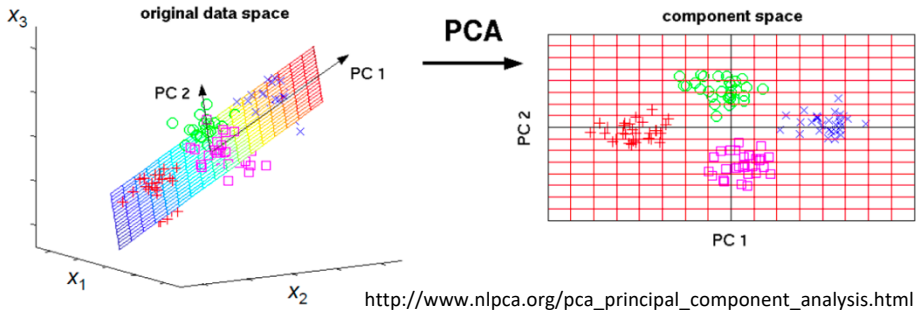


Figure II.11 Representation of the PCA method with the transformation of a 3-dimension space in two principal components

Obviously, we lost information in the process but the ratio of lost information is negligible if we chose the number of principal components as 25.

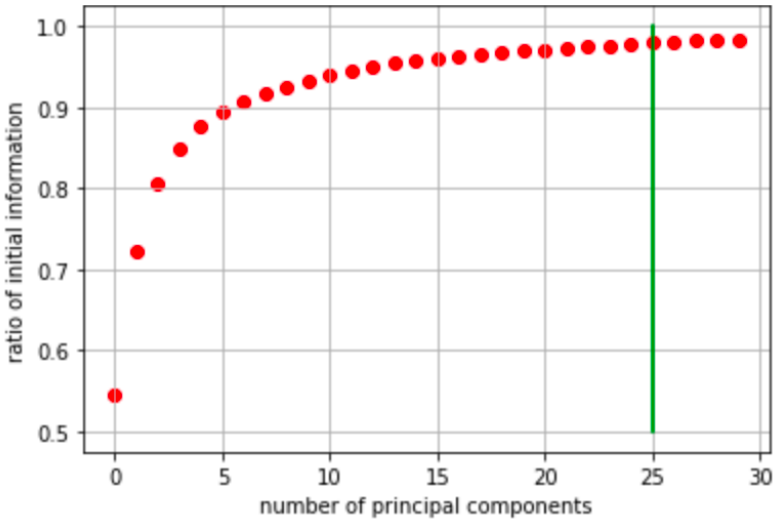


Figure II.12 Ratio of initial information kept after the PCA

We firstly tried to use the PCA as a classifier but the result were not satisfying. Thus, we chose to use the PCA as a feature by calculating the total mean error of the difference between each point of cardiac frequency before and after the PCA treatment. It allowed us to make a proper reconstitution of thirty minutes of an attack signal with only two minutes of surgery.

We defined the total error by:

$$\epsilon_{total} = \frac{1}{m} \sum_{j=1}^{m} ||CF^{j}_{aPCA} - CF^{j}_{bPCA}||^{2}_{2}$$

Where

- m : Number of frequency points for a patient signal
- CF_{bPCA}: Cardiac Frequency point before the PCA treatment
- CF_{aPCA} : Cardiac Frequency point after the PCA treatment
- $||.||_{2:}$ Euclidean norm

Our problem is that we did not know if we had a track of information in our data. The PCA reconstructs properly an attack signal, proof of the information into the signal.

II.2.d. Analysis Algorithms

II.2.d.i. Classifiers

Random Forest Algorithms:

A decisional tree is a mathematical object made of branches, nods, and leaves. Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value. The condition, or test, is represented as a nod and the possible outcomes as "leaves". This splitting process continues until no further gain can be made or a preset rule is met if the maximum depth of the tree is reached for instance. It then associates a given input to a predicted output by following the branches as in the example below.

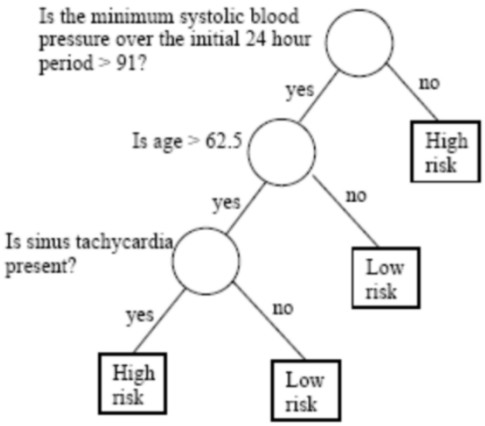


Figure II.13 Example of a decisional tree

Each nod corresponds to a test, and by following the branch associated with the output value of each test, the algorithm can associate each patient with a certain level of risk. Using a training and labelled dataset allows us to algorithmically build such a decisional tree. Buy randomly splitting the training dataset we can build several decisional trees which will, for each of them, predict an output label for each patient. The label that is most represented by the set of trees will finally be the output of the Random Forest. To build such decisional trees we apply a recursive algorithm to build trees that will minimize a given weight function.

Such an algorithm is very useful first because it is very efficient and also because it is understandable as it can return the features it mostly uses to classify the dataset.

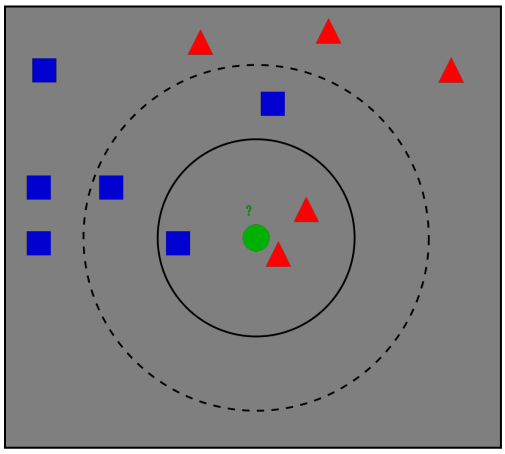
K-Nearest-Neighbors:

The K-Nearest Neighbor algorithm is based in one of the oldest classifying methods. It relies on the representation of studied objects, here patients with N corresponding features, as vectors in the N-dimensional space of the features. Assuming objects of the same class share the same characteristics, an object will then be classified in regard of the K objects that are the closest to it. The distance between objects is usually the Euclidean norm, and K is set by the user.[12]

For instance, a 1-NN classifier will label a patient the same as the nearest patient. A K-NN classifier will chose the label that appears the most among the K neighbors.

This algorithm may show its limits when one of the classes is overrepresented. It is also quite sensible to irrelevant features.

The main parameters of the algorithm are the number of neighbors (K), and the weight of the different neighbors, which can be uniform or inversely proportional to the distance to the neighbor.



- a 3-NN classifier would choose the red class

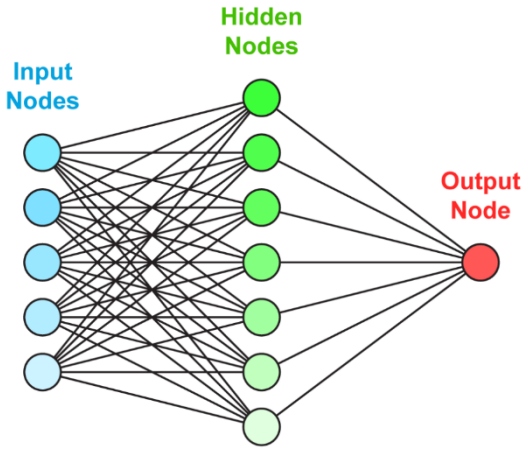
- a 5-NN classifier would choose the blue class

- a weighted 5-NN classifier would choose the red class.

Since the prediction relies on the distribution of objects in a N-dimensional space, the data has to be normalized before the classification.

Multi-Layer-Perceptron:

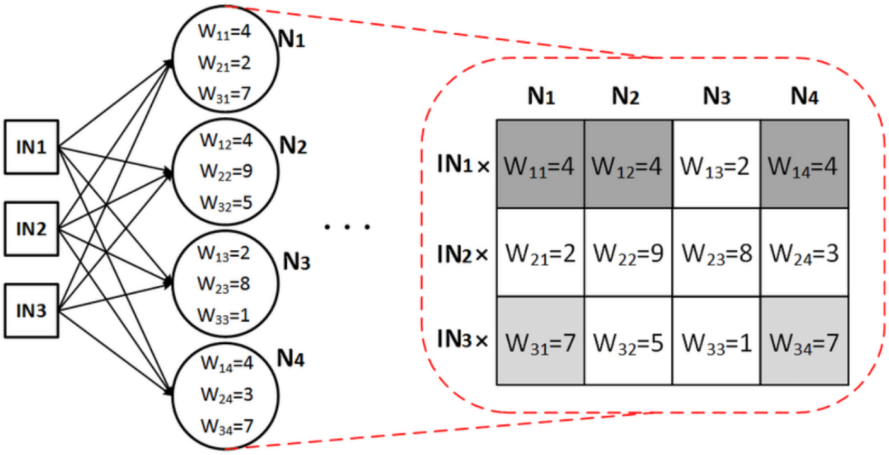
Neural networks are models inspired by the functioning cerebral cortex of human beings. It tries to replicate the same style of thought processing. They are organized in layers : -the input layer consists of a set of nodes representing the input features -the hidden layers, that have to intervene in data transfer between the input and output layer - the output layer : in our case, it consists of one node, which gives a number between 0 and 1. Depending on the threshold, this number is converted to a label : « clean » or « attack » Each node (corresponding to a neuron) is connected to all the nodes from the previous layer, and contains an activation function computing the output of the node. Each data coming from a previous node receives a weight and is multiplied and added. A bias is added and then passed to the activation function.[10]



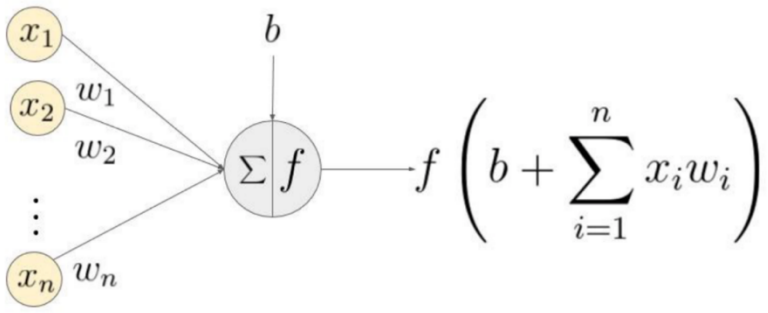
www.intellisystem.it/en/category/news/news-tecnologia/

Figure II.14 A Multilayer Perceptron

During the training phase, a training dataset is provided to the MLP. As shown below, it adjusts the weights and the bias, in order to recognize the characteristic features. This adjustment is made thanks to backpropagation. It consists of a gradient descent, where the loss function must be minimized. The perceptron finds the separating hyperplane that minimizes the number of misclassified samples.



https://www.researchgate.net/figure/A-simple-neural-network-and-the-mapping-of-the-firsthidden-layer-onto-a-43-Weight_fig2_292077006



An example of a neuron showing the input (x_1 - x_n), their corresponding weights (w_1 - w_n), a bias (b) and the activation function f applied to the weighted sum of the inputs.

When it is later given new, unlabeld data, the neural network tries to classify it. If the output is wrong, the network learns from its mistakes and improves its performance rate, until it becomes completely accurate. But in order to reach that point, it has to be provided a lot of training data, around 5 000 minimum. It is far from being our case, which alters the performance of our MLP.

The MLP trains on two arrays : X, of size (n_patients, n_features) and Y, containing only the label for each patient. The training dataset is split into a a training dataset and a validation dataset, allowing the MLP to adjust the weights. After training, the model is supposed to predict labels for new samples from the test dataset.

The hyperparameters include :

-the activation function (we used SeLu for the first layers, and sigmoid for the output)

-the number of hidden layers, and the number of nodes they contain (we used 2 hidden layers, of 30 then 11 nodes)

-the loss function (we used binary cross entropy)

-the number of epoch (we used 30 epoch, with a batch size of 1)

-the validation split ratio (we used 0.1)

-the weight given to the « attack » samples (we had less of them so we had to give them a bigger weight). We gave them 5 times more weight than the « clean » samples.

-the optimizer (we used Adam, with a learning rate of 0.0001)

II.2.d.ii. Comparative analysis of classifiers

Various classifiers all have their assets and their flaws, making them complementary. The « No free lunch » theorem proves that there can't be a universally best classification algorithm.

Using one classifier rather than another depends on the context. Some factors that help deciding which one is best are the size of the training dataset, the number of features, the correlation between features and how likely they are to overfit.

The K-NN is robust to noise in training dataset, and effective in case of a large training dataset containing numbers only. It starts processing data only after it is given a test observation to classify. But the computation time is high, and we have to determine the type of distance used as well as the value of K. It works better to find similar cases.

The Random Forest can provide understandable explanation over the prediction, which is especially relevant in the context of surgery. Health practitioners have to be able to understand how the algorithm decides, in order to take a decision. It can handle high dimensional spaces and works better to classify examples. Random Forest works well with a mixture of numerical and categorical features. It requires less preprocessing, and the training process is simpler. However, it is prone to outliers and tends to overfit if tree pruning is not used. It also doesn't perform well when some features depend on other features (this makes the trees of the forest less independent from each other).

MLP (Multi-layer perceptron) is a type of neural network. It outperforms Random Forest when there is sufficient training data. However, it doesn't provide any explanation over the decision, which make the features non interpretable. It is also more difficult to use and implement, considering the number of hyperparameters.

PCA (Primary Component Analysis) is the most popular dimensionality reduction algorithm. It identifies the hyper-plane closest to the data and projects the data onto it, by preserving as much information as possible. It reduces the number of features and gets rid of collinear features. It can be used as a classifier, but it is best to use it to reduce the number of features before giving the dataset to other classifiers such as the Random Forest.

II.2.d.iii. Finding the best classifier for our case

What criteria should our classifier maximise ?

From what we heard from the doctors, we had to priorities getting all the attacks, and not missing any, rather than prioritize accuracy and potentially missing attacks. The simple conclusion to this is that we would rather have more alarms than abnormal cases, as long as we get all the problematic cases. Indeed, we must keep in mind that our algorithm is a tool to help doctors, not to replace them, so providing the team with a new alert signal seems more coherent. We will then have to try to maximize the Recall variable, even if it means losing a little bit of Attack Accuracy, variables that we will study in the next paragraphs.

Prediction time:

The prediction time appeared to be of primary importance in all the interactions we had with doctors on this subject. Many choices may be different depending on this parameter alone: for example, if the prediction time is less than one minute, the anesthetists will not have the time to anticipate and alleviate the problem.

The recall variable: it represents the sensitivity of the algorithm, that is to say

$$Recall = P(Attack_{Predicted} \mid Attack)$$

As stated in the short introductory paragraph, the Recall variable must be maximized in our case.

Attack accuracy variable : it represents the positive predictive value of the algorithm, that is to say :

$$Precision_{Attack} = P(Attack \mid Attack_{Predicted})$$

As stated in the short introductory paragraph, the Attack Accuracy variable must be maximised, but is not as important as the Recall variable, so it can potentially be lowered in favour of an increase in the Recall variable. As a result, the best classifier for our case should respect such criteria.

The GridSearch Method

The step following feature engineering was finding an effective classifier. That meant running through all the combinations of classifiers, classifier parameters and features, and recording their performance: accuracy, recall, precision, and f1-score. Such an algorithm being extremely time-consuming, we decided to regroup features that we thought were sides of the same information. For instance, all the features that relied on the Fourier transform were always used simultaneously.The results were printed on a csv sheet showing for each classifier the features that were used and its score.Here is what the output looks like for a few features:

	classifier	energy_1	energy_2	energy_3	energy_4	pca	accuracy	recall_clean	recall_attack	precision_clean	precision_attack	f1_score_clean	f1_score_attack
0	RandomForestClassifier(class_weight='balanced', max_depth=5)						0.5988372093023255	0.6506849315068494	0.42857142857142855	0.8715596330275229	0.13953488372093023	0.7450980392156864	0.2105263157894737
0	RandomForestClassifier(class_weight='balanced', max_depth=5)						0.6686046511627907	0.6948051948051948	0.7142857142857143	0.963963963963964	0.10204081632653061	0.8075471698113208	0.17857142857142858
0	RandomForestClassifier(class_weight='balanced', max_depth=5)	X	X	X	X		0.7441860465116279	0.821917808219178	0.3333333333333333	0.9022556390977443	0.18518518518518517	0.8602150537634409	0.23809523809523808
0	RandomForestClassifier(class_weight='balanced', max_depth=5)					X	0.6627906976744186	0.7105263157894737	0.4166666666666667	0.9310344827586207	0.13513513513513514	0.8059701492537312	0.20408163265306126
0	RandomForestClassifier(class_weight='balanced', max_depth=5)						0.622093023255814	0.6778523489932886	0.2307692307692308	0.926605504587156	0.06818181818181818	0.7829457364341086	0.10526315789473682
0	RandomForestClassifier(class_weight='balanced', max_depth=5)	X	X	X	X		0.813953488372093	0.8979591836734694	0.5	0.9103448275862069	0.2727272727272727	0.9041095890410958	0.3529411764705882
0	RandomForestClassifier(class_weight='balanced', max_depth=5)					X	0.6453488372093024	0.6887417218543046	0.3333333333333333	0.9043478260869565	0.1	0.7819548872180451	0.15384615384615383
0	RandomForestClassifier(class_weight='balanced', max_depth=5)	X	X	X	X		0.75	0.7933333333333333	0.5714285714285714	0.937007874015748	0.24242424242424243	0.8592057761732852	0.3404255319148936
0	RandomForestClassifier(class_weight='balanced', max_depth=5)					X	0.7383720930232558	0.815068493150685	0.23529411764705882	0.9083969465648855	0.16666666666666666	0.8592057761732852	0.19512195121951217
0	RandomForestClassifier(class_weight='balanced', max_depth=5)	X	X	X	X	X	0.7790697674418605	0.8496732026143791	0.4444444444444444	0.948905109489051	0.13333333333333333	0.896551724137931	0.20512820512820512
0	RandomForestClassifier(class_weight='balanced', max_depth=5)	X	X	X	X		0.7790697674418605	0.8129032258064516	0.5	0.9545454545454546	0.18181818181818182	0.8780487804878049	0.26666666666666666
0	RandomForestClassifier(class_weight='balanced', max_depth=5)					X	0.622093023255814	0.6776315789473685	0.16666666666666666	0.9196428571428571	0.06060606060606061	0.7803030303030303	0.08888888888888888

The information gathered by the other team led us to choose the classifier having the highest recall and a decent precision, which was a RandomForest algorithm with these characteristics :

Classifier: Random Forest

Parameters: class_weight = 'balanced', max_depth = 5

Used features: mean_pouls,std_pouls, mean_SpO2, std_SpO2, energy_1, energy_2, energy_3, energy_4, moment3_Pouls, moment4_Pouls, moment3_SpO2, moment4_SpO2

III. Results and limits

III.1. Understanding of the results

In order to test the algorithm, we tried to make it predict the label between Anomaly or Attack (in the same category) and Clean. The dataset used for the test contains the operations of the month of November, it is a new dataset that the algorithm had never seen. Our approach is quite ambitious as in the operating room, data analysis is not always considered as a possibly useful tool, especially by surgeons. Furthermore, the data we studied consists of only two parameters sampled every 5 seconds, which is extremely poor. Any relevant result would then be an important step forward.

Recall	Precision
0,41	0,44

These results were found while trying to maximize recall, however the precision value is relatively satisfying too. While they might appear insufficient, they must be considered qualitatively. Indeed, the performance of the algorithm is better than what a random system would do if it considered the proportion of attacks among all the operations. Moreover, for better understanding of the results, we looked at the false-positive cases.

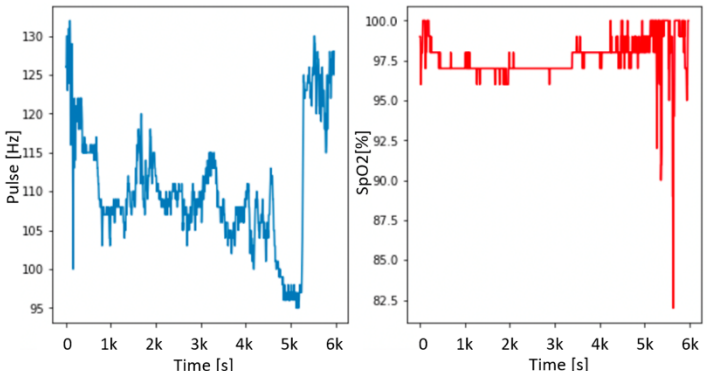


Figure III.1 False positive data

It showed that these cases are ambiguous ones that even human struggle to classify. It means that there is undeniably some relevant information to be studied, which has not always been obvious. The algorithm is interpretable since it provides the main features on which it is based when it operates. However, not all of these characteristics are currently visible to a doctor on a scope. Moreover, our algorithm uses only the first thirty minutes of the operation, unlike the medical team, which is more reactive and will necessarily remain alert throughout the operation. We observe a difference between the reasoning of our tool and human reasoning, which is very interesting since we place ourselves in a predictive perspective which differs from medical empiricism.

A solution that stems from the lack of data and this new perspective is the subsequent use in knowledge building regarding accidents in operating rooms. Indeed, if the characteristics are poor, the final solution is likely to stay outside of the operating room and only be used for educational purposes. Hence, the data collected by our algorithm will

be used in a second phase to potentially respond to misunderstandings following an anomaly in an operating room, or simply to try to understand the method used by the algorithm, with the aim of reproducing it during operations. That way, by looking at a more predictive lens through our algorithm data, we might change the paradigms of medicine.

The fact that poor data gave such results is extremely encouraging for the future of this approach. Indeed, if data collecting is improved, the results might become excellent. Consequently, it is possible to imagine a situation where the algorithm is extremely performant to predict cardiac attacks. In this case, we submitted a model of the utility of the algorithm depending on its performances (cf appendix.2)

However, for the algorithm to be efficiently used and to prove its relevance, we need to understand how anesthetists will include the algorithm in their decision process.

III.2. Algorithm and decision making

The prediction algorithm is supposed to be used by the anesthetist who is responsible for the patient's safety during the time of the operation. They are responsible for the induction and awaking phase. They control the anesthesia with the artificial respirator, and medicine delivered through perfusions. They spend time watching the patient's ECG, oxygen saturation, arterial pressure, acapnia, respiratory frequency, but also the patient's appearance, especially the way they breath and the color of their face.

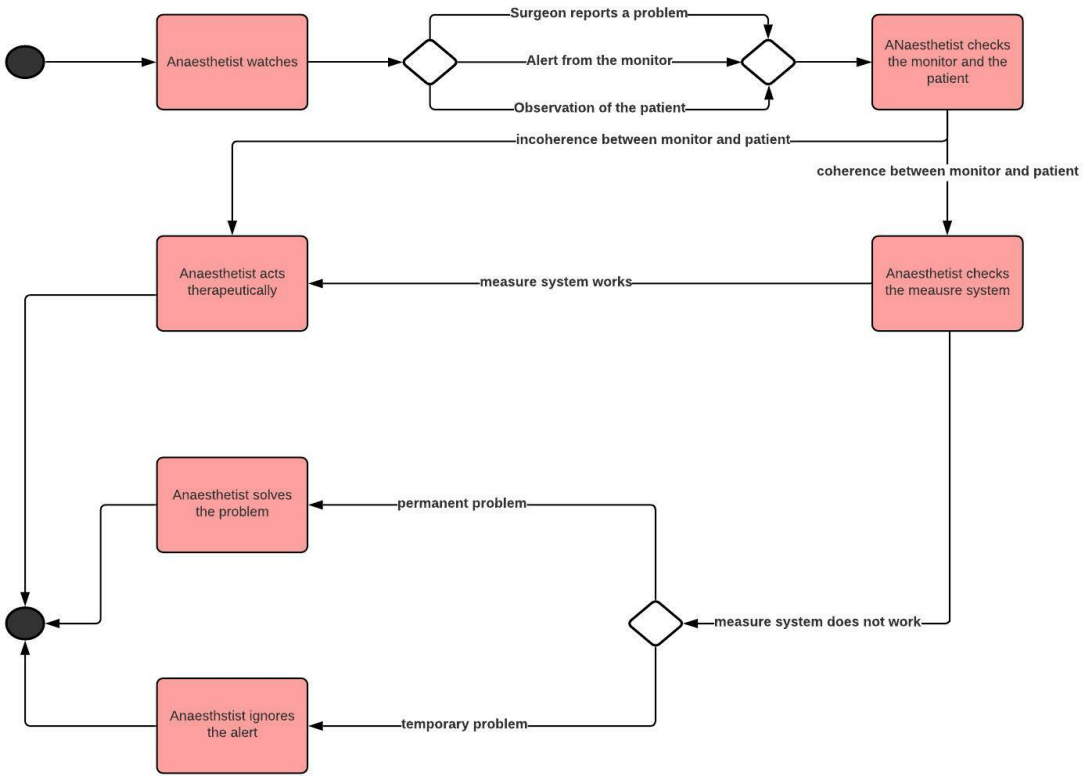


Figure III.2 Decision making in the operating room without the algorithm

Our observation showed that the anesthetists are the ones who receive information about the patient and know how to deal with contradictory information and how to react. The algorithm should then have the same role in the room as the data from the monitor, alerting the anesthetist by collecting data from the patient.

However, the decisional process of the anesthetist is supposed to be modified by the system. We imagined three possible reactions from the anesthetist: indifference, anticipation, and action.

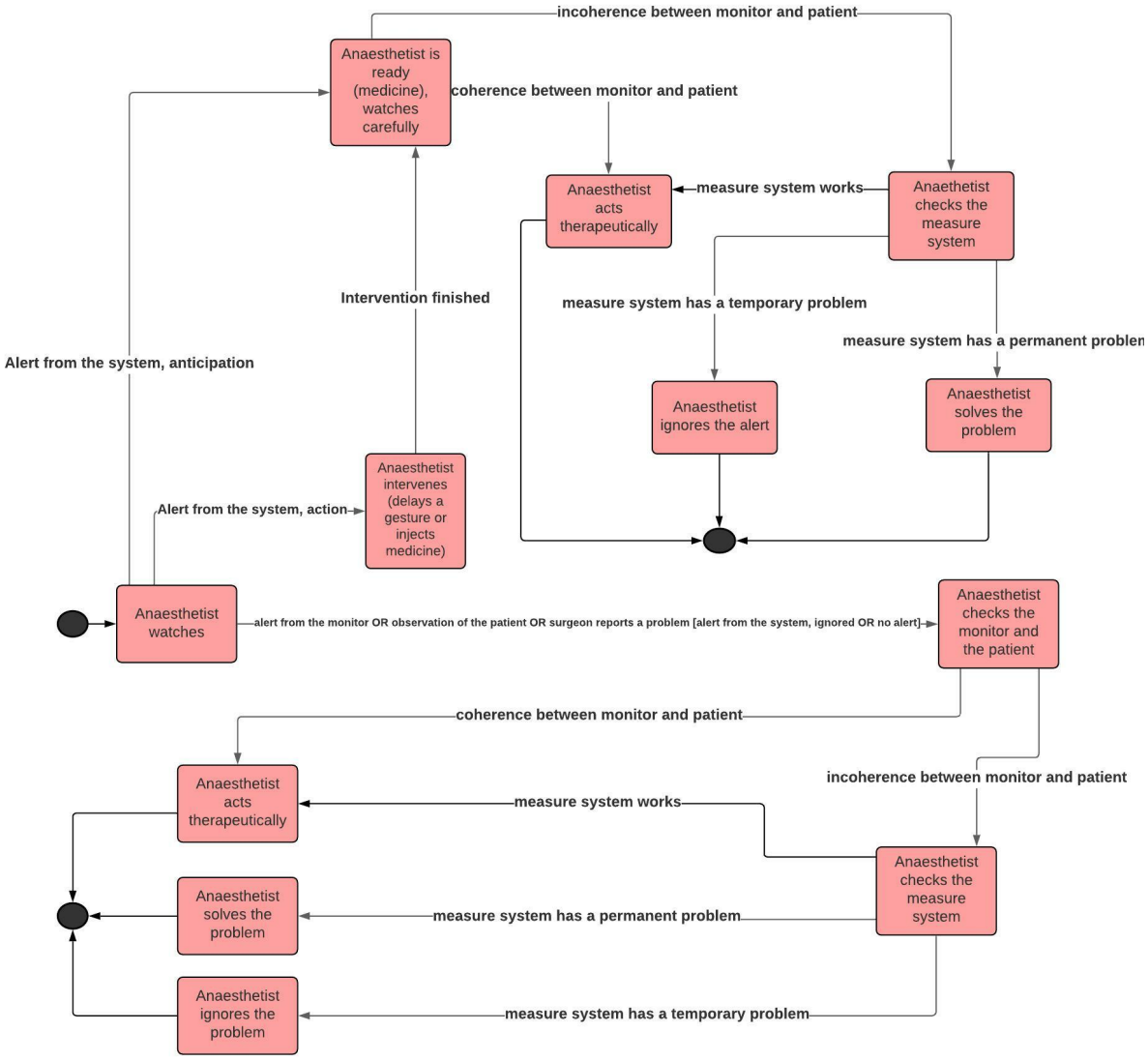


Figure III.2 Decision making in the operating room with the algorithm

IV. Conclusion

The Ethics Guideline on trustworthy AI[1] by the independent high-level expert group on AI, came up in 2019 with 4 principles that a trustworthy AI must respect:

Respect of human autonomy

Prevention of harm

Fairness

Explicability

A quick thought on engineering, tells us that failure is an integral part of engineering and that the efficiency of a solution is to be considered according to the context. The context in which we are working is unfriendly: data is lacking, noisy and there are multiple pathologies. Now, knowing the defects the final solution might have, is it acceptable to influence anesthetists in their choice on such a critical type of decision?

First, we learned that the anesthetists will not immediately inject medicine to the patient after an alert but rather get prepared. There is no risk for the patient to be directly impacted by a mistake from the algorithm: the prevention of harm seems to be respected. Moreover, anesthetists are used to dealing with multiple monitors and alerts, they therefore have the capacity to integrate the information to decide, without being blinded by it. It can be useful to remind that the tool will be used for information purposes only, the idea is not to replace anesthetists but to assist and help them during the decision process. Thus, the respect of human autonomy is considered. As the algorithm might be able to tell which feature was used to make a decision, the issue of the explicability could be solved.

Apart from that, the impact of the algorithm is still not to be undermined as it could catch anesthetists' attention for no reason, and because repetitions of alerts might diminish their compliance with the tool. Another adverse effect is the fact that the compliance on the tool might be too great that when it misses a problem, anesthetists diminish their concentration because there is no alert. The issue of the responsibility of the tool if something goes wrong is, then, still to be raised as it influences the practitioner in their choice. The creator of the tool might thus be held responsible. Moreover, as the algorithm is based on the quite subjective labelling made from only 3 curves (cardiac frequency, oxygen saturation and temperature) sometimes missing, the validity of the algorithm can be questioned.

VI. Appendix

Appendix.1 List of the surgeries we attended:

- hypospadias repair 1 (11/22 – OR 5 – 3-year-old patient)
- hypospadias repair 2 (11/22 – OR 5 – 3-year-old patient)
- exploratory laparotomy with two anastomoses (11/22 – OR 5 – 5-month-old patient)
- central line insertion (11/22 – OR 6 – 9-year-old patient)
- chest drainage (11/22 – OR 6 – 8-year-old patient)
- heart surgery (11/23 – OR 4 – 4-year-old patient)
- ablation of osteosynthesis material (11/23 – OR 1 – 15-year-old patient)
- ablation of osteosynthesis material (11/23 – OR 2 – 29-year-old patient)
- pyloric stenosis repair (11/23 – OR 1 – 1-month-old patient)
- undescended testicle repair 1 (11/23 – OR 5 – 1 year-old-patient)
- undescended testicle repair 2 (11/23 – OR 5 – 8 year-old-patient)
- inguinal hernia repair (11/23 – OR 5 – 6-month-old patient)
- male circumcision 1 (11/23 – OR 5 – 5-year-old patient)
- bronchoalveolar lavage 1 (11/24 – OR 8 – 3-year-old patient)
- bronchoalveolar lavage 2 (11/24 – OR 8 – 15-year-old patient)
- esophageal dilation (11/24 – OR 8 – 2-year-old patient)
- cleft lip repair (11/24 – OR 5 – 2-month-old patient)
- JJ stent insertion (11/24 – OR 6 – 9-year-old patient)
- fiberoptic endoscopy with airway management (11/24 – OR 8 – 2-month old patient)
- aortic coarctation repair (11/25 – OR 4 – 14-day-old patient)
- male circumcision 2 (11/25 – OR 6 – 3-year-old patient)
- cleft palate repair (11/25 – OR 6 – 6-month-old patient)

Appendix 2 Scheme of the medical reaction according to what the algorithm responds

Here a distinction is made between several levels of recall and anticipation to consider what the practitioner's reaction might be. At the end of the day, it turns out that less than a minute's anticipation is completely useless during surgery because the practitioner will not have enough time to anticipate. Similarly, with a 50% recall, the confidence level might be too low to be considered a serious threat. The confidence level increases with recall, and the quality of the reaction with the time to anticipate. However, a line between a useful and a useless tool according to these two characteristics is difficult to draw and really depends on the anaesthetist's level of confidence in the tool.

Thus, it is necessary to specify the two main possible uses of our tool in this context :

The first is obviously the one requested by the customer; active use in an operating room to predict accidents.

However, there is another underlying solution that stems from a reflection on the lack of data, the subsequent use in research and knowledge building. Indeed, if the characteristics are poor, the final solution is likely to stay outside of the operating room and only be used for educational purposes.

Throughout our journey through the operating theatres, we discovered skepticism among health professionals towards new technologies, especially for old-style practitioners who emphasize clinical signs over monitor signs. As a result, the real challenge for the tool could be to be accepted and used by practitioners.[5]

For our scheme, it is necessary to specify the quantities that are used in order to discern the different cases:

T is the prediction time of our algorithm.

The degree of interpretation, which takes the values 0, 1 and 2. Degree 0 is equivalent to a silent algorithm; giving no explanation. Degree 1 is equivalent to an algorithm that returns partial pieces of feature, or whole but obscure features but may seem obscure to an uninformed physician (typically: 6th coefficient of the Fourier series decomposition of SpO2). Finally, degree 2 is equivalent to an algorithm that delivers simple and accurate information about what it has been based on, allowing doctors to react in an efficient way.

VII. Bibliography

1. Anonymous. « Ethics Guidelines for Trustworthy AI ». Text. Shaping Europe's digital future - European Commission, 8 avril 2019. https://wayback.archive-it.org/12090/20201227221227/https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
2. Azencott, Chloé-Agathe. *Introduction au machine learning*. Dunod, 2019.
3. Bertsimas, Dimitris, Luca Mingardi, et Bartolomeo Stellato. « Machine Learning for Real-Time Heart Disease Prediction ». *IEEE Journal of Biomedical and Health Informatics*, 2021.
4. Breiman, Leo. « Bagging predictors ». *Machine learning* 24, nº 2 (1996): 123-40.
5. Callon, Michel. « Éléments pour une sociologie de la traduction: la domestication des coquilles Saint-Jacques et des marins-pêcheurs dans la baie de Saint-Brieuc ». *L'Année sociologique (1940/1948-)* 36 (1986): 169-208.
6. Calvino-Casilda, Vanesa, Antonio José López-Peinado, Rosa María Martín-Aranda, et Elena Pérez-Mayoral. « Applications and Technologies », 2019.
7. Duffau, Hugues. *L'erreur de Broca*. Michel Lafon, 2016.
8. Gañán-Cárdenas, Eduard Alexander, Jorge Isaac Pemberthy-Ruiz, Juan Carlos Rivera-Agudelo, et Maria Clara Mendoza-Arango. « Operating Room Time Prediction: An Application of Latent Class Analysis and Machine Learning ». *Ingenieria y Universidad* 26 (2022).
9. Mallet, Y., O. De Vel, et D. Coomans. « Fundamentals of wavelet transformations ». édité par Beata Walczak, 57-84. Amsterdam, The Netherlands: Elsevier, 2000. http://www.elsevier.com/books/wavelets-in-chemistry/walczak/978-0-444-50111-0.
10. Popescu, MARIUS-CONSTANTIN, VALENTINA Balas, ONISIFOR Olaru, et NIKOS Mastorakis. « The backpropagation algorithm functions for the multilayer perceptron ». In *Proceedings of the 11th WSEAS International Conference on Sustainability in Science Engineering*, 28-31, 2009.
11. Richardson, M. « Fundamentals of the discrete Fourier transform ». *Sound & Vibration Magazine*, 1978, 1-8.
12. Cover, T., et P. Hart. « Nearest Neighbor Pattern Classification ». IEEE Transactions on Information Theory 13, nº1 (janvier 1967): 21-27. https://doi.org/10.1109/TIT.1967.1053964

VIII. Special thanks

Xavier Alacoque

Sebastien Travadel

Franck Guarnieri

Didier Delaitre

Mathis Bourdin

University Hospital Center of Toulouse

Oncology Resarch Center of Toulouse

CRC Mines Paris