Highlights

ICP-SWAn: A Selective Waveform Analysis Pipeline for Conformal Subpeak Detection in Intracranial Pressure Signals

D. Legé, B. Balança, A. Kazimierska, G. Percevault, V. Ghibaudo, C. Kokonendji, M. Prud'homme, J. Henriet

- In an online setting, P2/P1 ratio calculation should include proper quality control.
- Certain precautions must be taken when performing real-time ICP signal analysis.
- The benefits of conformal prediction are mostly seeable in a noisy environment.

ICP-SWAn: A Selective Waveform Analysis Pipeline for Conformal Subpeak Detection in Intracranial Pressure Signals

D. Legé[a,b], B. Balança[c], A. Kazimierska[d], G. Percevault[c], V. Ghibaudo[c], C. Kokonendji[e], M. Prud'homme[b], J. Henriet[a]

[a] *Université de Franche-Comté, CNRS, Institut FEMTO-ST, 16 route de Gray, Besançon 25000, France*
[b] *Sophysa, 5 rue Guy Môquet, 91400 Orsay, France*
[c] *Department of Neurological Anesthesiology and Intensive Care Medicine, Hospices Civils de Lyon, Hôpital Pierre Wertheimer, 59 boulevard Pinel, Bron 69500, France*
[d] *Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, 27 Wybrzeze Wyspianskiego Street, Wroclaw 50-370, Poland*
[e] *Université Marie et Louis Pasteur, Laboratoire de Mathématiques de Besançon – UMR 6623 CNRS-UMLP, 16 route de Gray, Besançon 25000, France*

Abstract

Intensive care unit management of patients with traumatic brain injury or aneurysmal hemorrhage often requires intracranial pressure (ICP) monitoring. Although the primary goal is to prevent adverse intracranial hypertensive events, mechanical properties of the cerebrospinal system can also be derived from a closer analysis of the ICP signal. The morphology of the ICP signal can be influenced by a variety of physiological components and is therefore highly variable. In particular, on the time scale of a heartbeat, characteristic subpeaks called P1, P2, and P3 are visible most of the time and their analysis may provide valuable insights about cerebrospinal pressure-volume compensation. However, their automatic detection in real time is a challenging task that requires displaying no result in situations of uncertainty to avoid misleading the clinicians. Therefore, we applied several results from the "Learn-then-Test" theoretical framework to our deep learning-based delineation algorithm to identify cases where it is appropriate to refrain from providing a result. The subpeak detection task was performed using recurrent neural networks trained on a 59-patient dataset. Performance was evaluated on two independent test datasets. The first one consisted of 13,086 pulses extracted from 49 recordings and labeled by 3 independent operators. The second one, which is publicly available, consisted of 11,172 pulses extracted from 63 recordings. The mean absolute error on the calculated P2/P1 ratio was 0.061 for the first test dataset and 0.016 for the second test dataset.

Keywords: Intracranial Pressure, Intracranial Compliance, Conformal Prediction, Recurrent Neural Networks, LSTM

*This work has been done in the frame of EIPHI Graduate School (contract ANR-17-EURE-0002) and supported by Bourgogne-Franche-Comté Region, the French Research Ministry, CNRS (French National Center for Scientific Research), and Sophysa Company (Orsay, France).

*Corresponding author
Email address: donatien.lege@femto-st.fr (D. Legé)

1. Introduction

Intracranial pressure (ICP) monitoring is a key component of brain injury management in intensive care units (ICUs). The main focus for the clinician is to maintain mean ICP below a certain hypertension threshold, generally set to about 20 mmHg [1]. However, it is broadly recognized that the full ICP signal can provide valuable information about the physical properties of the cerebrospinal system that cannot be summed up by a simple rolling mean [2]. In fact, the ICP signal is a complex mixture of multiple physiological components, particularly affected by arterial and intrathoracic pressures [3]. When looking at cardiac-induced ICP pulses at the time scale of a single cardiac cycle, three subpeaks are generally visible (see Figure 1). These are commonly called P1, P2, and P3, in accordance with their order of appearance. The subpeak P1 is a percussion wave caused by the systolic blood influx into the cerebral vasculature [4]. The second one, P2, is a reflection wave that coincides with a maximum of blood volume increase in the cerebral arteries [5]. The origin of P3 remains unclear, but it tends to be associated with the dicrotic notch [6] and the venous outflow [7]. The relative heights of the subpeaks have been known to reflect the volume buffering capacity of the cerebrospinal space [8], often referred to as intracranial compliance (ICC). Under nonpathological conditions, P1 is superior to P2, which is superior to P3. As ICC decreases, the pulse shifts towards a triangular shape centered around P2, with P2 superior to P1 and P3. Therefore, the P2/P1 ratio, when measurable, can be used as surrogate for ICC. At the bedside, this information can help the clinician identify at-risk patients and adjust the therapeutic intensity [9][10].

However, real-time monitoring of the P2/P1 ratio is challenging for several reasons. Firstly, as with any medical recording, the ICP signal is subject to artifact pollution [11]. These artifacts are mainly due to electronic noise, coughing, and patient handling. Secondly, due to its numerous physiological dependencies, the ICP waveform is highly variable between patients. Finally, subpeaks P1, P2, and P3 are not always visible. For instance, a previous study reported that up to 10% of cardiac-induced pulses show at least one missing subpeak [12]. While several detection algorithms proposed in the literature attempted to tackle these uncertainty factors [13][14], none of them can offer statistical guarantees on the final performance in real-life signals. Although rarely mentioned in the literature, the interpretation of the ICP waveform can be subtle and subject to human interpretation, since the physiological mechanisms are not directly observable. Being able to refrain from giving a likely incorrect information is all the more suitable since the "alarm fatigue" in the intensive care environment is a well-documented phenomenon [15]. The display of a misleading P2/P1 ratio could be avoided by implementing a decision rule to identify uncertainty situations. In the present study, we introduce the version 1.0 of our signal analysis pipeline called ICP-SWAn (Selective Waveform Analysis). It integrates several results from the Learn-then-Test framework [16] into our previously published P2/P1 ratio calculation algorithm [17] to simultaneously control multiple risks at a user-defined level. After performing multiple hypothesis tests on a calibration dataset, simple thresholds could be associated with each neural network output to identify uncertainty regions where no P2/P1 value should be displayed. The final detection performance was assessed using a dataset publicly available at the following address: https://github.com/NeuroResearchCore/trackLight/tree/master/Data.

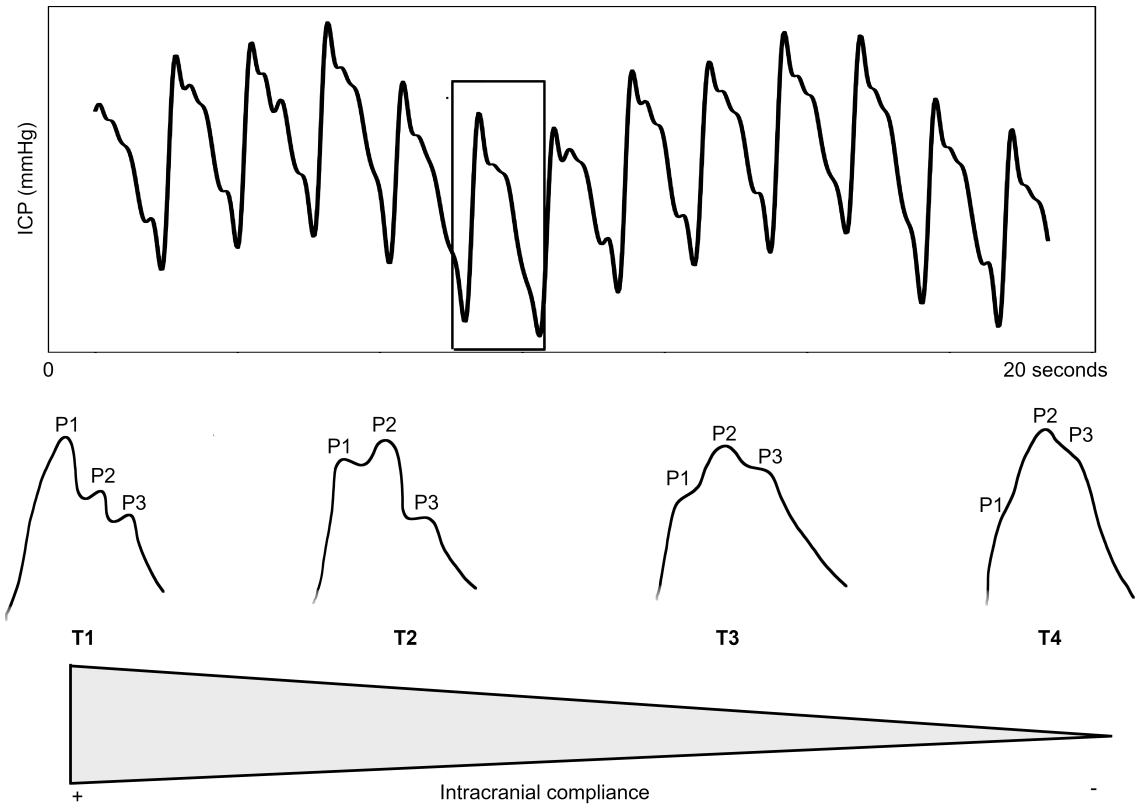

Figure 1: Positions of subpeaks P1, P2, and P3 in various types of intracranial pressure (ICP) pulse waveforms. The indication "Tn" corresponds to the associated pulse shape derived from an intelligence-based morphological classification presented in [18].

2. Material and Methods

2.1. Data Collection

ICP signals were acquired in two separated patient cohorts admitted to a neuro-ICU (NICU) between January 2021 and March 2024. One cohort (49 patients) was dedicated to the training process and the other (59 patients) to the testing process. In both cases, patients were treated according to the international guidelines for the management of acute brain injuries [19]. ICP signals were acquired with an intraparenchymal sensor (Pressio, Sophysa, Orsay, France) at a sampling rate of 100 Hz. In the case of the test dataset, a synchronized 100-Hz arterial blood pressure (ABP) signal was also acquired using an invasive arterial catheter. Inclusion criteria were as follows: age between 18 and 80 years and sedation under mechanical ventilation. Exclusion criteria included major hemodynamic failure and decompressive craniectomy. In addition, we used a public dataset to perform a second test process. This dataset, associated with a comparative study on the automated detection of ICP subpeaks [20], comprised 11,172 valid pulses extracted from 64 patient recordings acquired with an intraparenchymal sensor with a minimum 240 Hz resolution. The final data partitioning is shown in Figure 2. The use of two different test datasets is a way of highlighting the performance variations that can be induced by the evaluation context. Although the content of the public dataset is excellent for assessing the detection performance of P2/P1 *in a vacuum*, our custom dataset was designed to cover most of the difficulties that can occur in a real-time setting by including segments with artifacts (see in Table 1).

Regarding the composition of the test cohort, the SAH patient population was 40% male with a mean age of 56.4 years. The TBI population was 88% male with a mean age of 44.6 years, peaking between 30 and 40 years. In both cases, these proportions are generally representative of the epidemiology observed in western Europe [21][22], although with a slight overrepresentation of male patients in the TBI population.

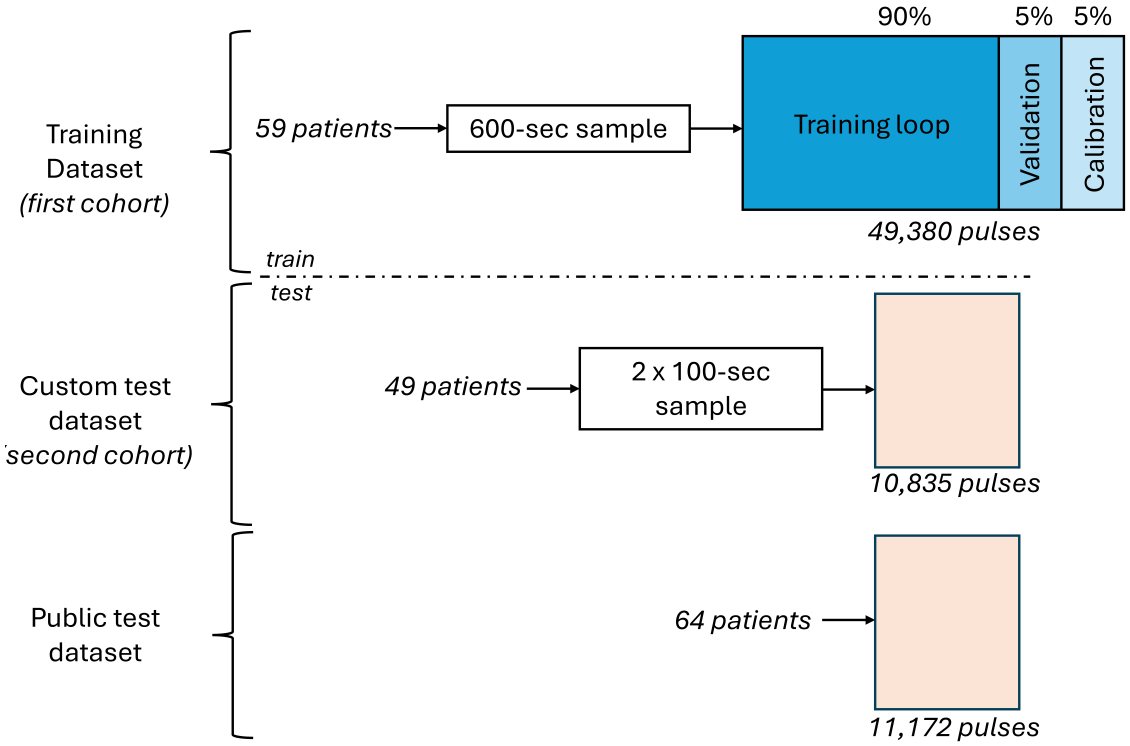

Figure 2: Data Partitioning

2.2. Annotation Process

The researchers in charge of the labeling process used a custom-made Python annotating tool and had access to the entire ICP recordings. In the case of the test dataset, they also had access to a synchronized auxiliary ABP signal as an interpretation aid. To build the training set, one 600-second segment was manually extracted from each of the 59 recordings included in the training test. To build the test set, two 100-second segments were manually extracted from each of the 49 included records, with a minimum delay of 24 hours, creating a total of 98 ICP segments. This extraction process was applied to cover a wide range of waveforms based on three criteria: (i) mean ICP values, (ii) P2/P1 ratio, and (iii) the overall waveform morphology as assessed by an AI-based classification [18]; see Table 1. For each individual pulse in these segments, P1 and P2 were manually annotated when identifiable. Otherwise, the pulse was classified either as non-interpretable or as an artifact. The training dataset was entirely annotated by the first author (DL). In contrast, the annotation process for the test dataset was designed as follows: (i) 94 ICP segments were evenly distributed among three annotators (VG, GP, or AK), with each segment assigned to a single author. (ii) The remaining 4 samples (accounting for 514 pulses) were labeled by each of the three annotators to validate the inter-annotator agreement. Independently of this procedure, 10 recordings were reported as particularly noisy and difficult to interpret during the annotation

process. Consequently, these records were re-labeled by each of the three annotators in order to adopt the majority verdict.

Demographic data	Training dataset	Custom test dataset	Public test dataset
n	59	49	64
Age	53, std = 17	49, std = 12	-
Sex	19 (F) / 40 (M)	21 (F) / 28 (M)	-
TBI	42	28	-
SAH	11	15	-
Other pathology	6	6	-
ICP segment statistics			
Mean ICP (mmHg)	11.90, std = 9.7	8.88, std = 8.11	3.46, std = 7.28
Mean P2/P1 ratio	1.22, std = 0.59	1.44, std = 0.58	1.25, std = 0.34
Full dataset statistics			
Valid Pulses	49,380 (86.4%)	10,835 (82.7%)	11,172 (100%)
Non-Interpretable Pulses	3,988 (6.9%)	838 (6.4%)	-
Artifacts	3,889 (6.7%)	1,413 (10.8%)	-
Pulse Shape Class			
T1 - Normal	15,655 (27.7%)	2,151 (16.4%)	1,486 (13.3%)
T2 - Possibly Pathological	17,041 (30.2%)	3,232 (24.7%)	5,994 (53.7%)
T3 - Likely Pathological	9,554 (16.9%)	4,392 (33.6%)	2,811 (25.2%)
T4 - Pathological	5,507 (9.74%)	1,692 (12.9%)	514 (4.6%)
A+E - Artifacts / Errors	8,758 (15.5%)	1,619 (12.4%)	367 (3.3%)
Total	57,257	13,086	11,172

Table 1: Dataset composition. The pulse shape class corresponds to the output of a deep-learning based pulse waveform classification and is only used here as an indicator of the waveform diversity. The "A+E" class includes more cases than the artifacts which where manually labeled, which explains the differences between the two percentages. n — number of patients, std — standard deviation, TBI — traumatic brain injury, SAH — subarachnoidal hemorrhage, ICP — intracranial pressure

2.3. Preprocessing

Prior to analyses, a fourth-order low-pass Butterworth filter with a cutoff set to 20 Hz was applied to all of the recordings to remove high frequency noise. Cardiac pulses were isolated using the modified Scholkmann algorithm [23]. Each pulse extracted from the filtered ICP signals was preprocessed as follows:

1. A third-degree polynomial interpolation was applied to set the pulse length to 180 points and remove differences in pulse duration stemming from changes in heart rate.
2. The straight line passing through the pulse onset point and the pulse end point was subtracted from the 180-point vector to remove the vertical misalignment of the pulse stemming from the respiratory component of ICP.
3. The pulse height was normalized between 0 and 1 to remove differences in pulse amplitude.

Pulses extracted from the training cohort were randomly divided into three subsets: (i) The first one (90%) was used to perform the actual training loop. (ii) The second one (5%) was used as a validation subset. (iii) The last one (5%), also called the *calibration* dataset,

was dedicated to the multiple risk control procedure. This separation was done so that all the pulses collected from a given patient were exclusively dedicated to either the actual training loop or the validation/calibration steps.

2.4. Data Augmentation

Synthetic examples were generated to increase the variability of the subset dedicated to the training process. The data enhancement procedure consisted of adding multivariate Gaussian noise to the first principal components of locally average pulses. More formally, the procedure was as follows:

1. A principal component analysis (PCA) was performed on the real-world dataset described earlier.
2. For each example x_i, the n nearest neighbors according to the Euclidean distance calculated on the p first principal components are identified. A $p \times p$ covariance matrix Σ is calculated based on these $n + 1$ selected examples.
3. The barycenter of the $n + 1$ selected examples is calculated. A multivariate Gaussian noise is added to it. The parameters of the associated distribution are the following: The mean is set to 0 and the covariance matrix is set to Σ multiplied by an inflation factor k.
4. The generated point in the principal-component space is transformed back to the initial space. This synthetic pulse is preprocessed as described in Section 2.3.
5. The generated pulse is assigned to the same class as x_i. Where applicable, P1 and P2 are identified as the maxima of curvature that are the closest to the original P1 and P2 marked on x_i.

In practice, an entire new dataset was first generated. The parameters (n, p, k) were set to (5, 10, 10). A more targeted augmentation was then performed to increase the presence of artifacts in the training dataset. For this second procedure, the parameters (n, p, k) were set to (5, 10, 50). In the end, the augmented training dataset consisted in 178,930 pulses, among which 119,306 were valid pulses, 8000 were non-interpretable pulses, and 51,624 were artifacts. This procedure did not affect the validation dataset or the test datasets.

2.5. Subpeak Detection Pipeline

Certain modifications were introduced to the initial pipeline developed by the authors [17] to integrate methods of the Learn-then-Test framework. As described in the original paper, two neural networks successively process the input data. The first one is a LSTM-based classifier meant to eliminate artifacts and pulses without a calculable P2/P1 ratio. The second one is a Long Short-Term Memory (LSTM)-based subpeak detector designed to output a prediction interval for P1 and P2. Models were implemented using Pytorch 2.0 on Python 3.11.3. Experiments were run on a Windows 10 machine powered by WSL2 Ubuntu 20.04.5, equipped with a 12th Gen Intel(R) Core(TM) i7-12850HX 2.10 GHz 16 CPU, an Nvidia RTX A3000 12GB Laptop GPU, and 16 GB of RAM.

2.5.1. Pulse Selection

Firstly, preprocessed pulses are separated into three classes: (i) valid pulses, (ii) pulses without a calculable P2/P1 ratio, and (iii) artifacts (see in Figure 3). Although only the valid pulses are sent further to the detection step, we chose to distinguish three cases to facilitate

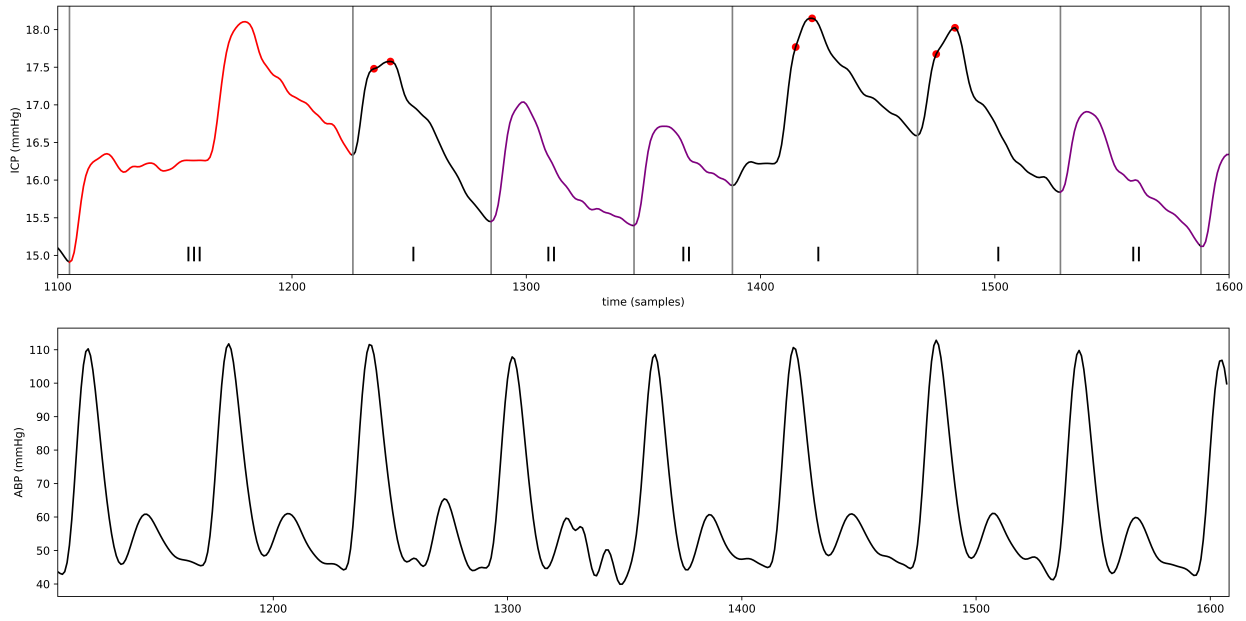

Figure 3: Examples of the three pulse classes considered in the P2/P1 calculation pipeline. Both signals were acquired with a sampling frequency of 100 Hz. ICP — intracranial pressure. ABP — arterial blood pressure. I — "valid pulse". II — "non-interpretable pulse". III — "artifact".

the annotation process and to improve the neural network convergence. In the experiments, we used a bidirectional LSTM network associated with a cross-entropy loss. The training process was carried out over 100 epochs with a batch size of 512.

2.5.2. P1 and P2 Designation

On valid pulses, a quantile regression is performed by a second neural network to output two prediction intervals around the positions of P1 and P2, respectively denoted as x_{P1} and x_{P2}. More formally, for a pulse x, the neural network outputs the following four-component vector: $(\ell_{p1}(x),\ u_{p1}(x),\ \ell_{p2}(x),\ u_{p2}(x))$. A prediction interval for x_{P1} is given by $[\ell_{p1}(x),\ u_{p1}(x)]$, whereas a prediction interval for x_{P2} is given by $[\ell_{p2}(x),\ u_{p2}(x)]$. Afterwards, we calculate the candidate set $\Gamma = (\gamma_i)_{i=1}^{n}$ that corresponds to the n positive local maxima of the vector $\kappa(-1000*x)$, with $\kappa(s) = s''/(1+s'^2)^{3/2}$ and where s' and s'' designate the first and second discrete derivatives of a pulse s, respectively. The function κ is generally called the *signed curvature* of a signal. The candidate selection procedure is inspired from the original MOCAIP algorithm [13]. Finally, x_{P1} and x_{P2} are estimated as follows:

$$\widehat{x}_{P1} = \underset{i\in[\![1,n]\!]}{\arg\min}[(\gamma_i - \ell_{p1}(x))^2 + (\gamma_i - u_{p1}(x))^2]$$

and

$$\widehat{x}_{P2} = \underset{j\in[\![P1,n]\!]}{\arg\min}[(\gamma_j - \ell_{p2}(x))^2 + (\gamma_j - u_{p2}(x))^2].$$

The final $\widehat{x}_{P1}$ and $\widehat{x}_{P2}$ are not necessarily included in their associated prediction intervals, as in the example in Figure 4. The aim of calculating prediction intervals is to obtain a bandwidth that can be used as an indicator of the model's uncertainty. In the experiments, we used a bidirectional LSTM network associated with two opposed 25%-pinball losses (i.e. two asymetric L1-losses), designed to output a 50%-prediction interval. The training process was carried out over 300 epochs with a batch size of 512.

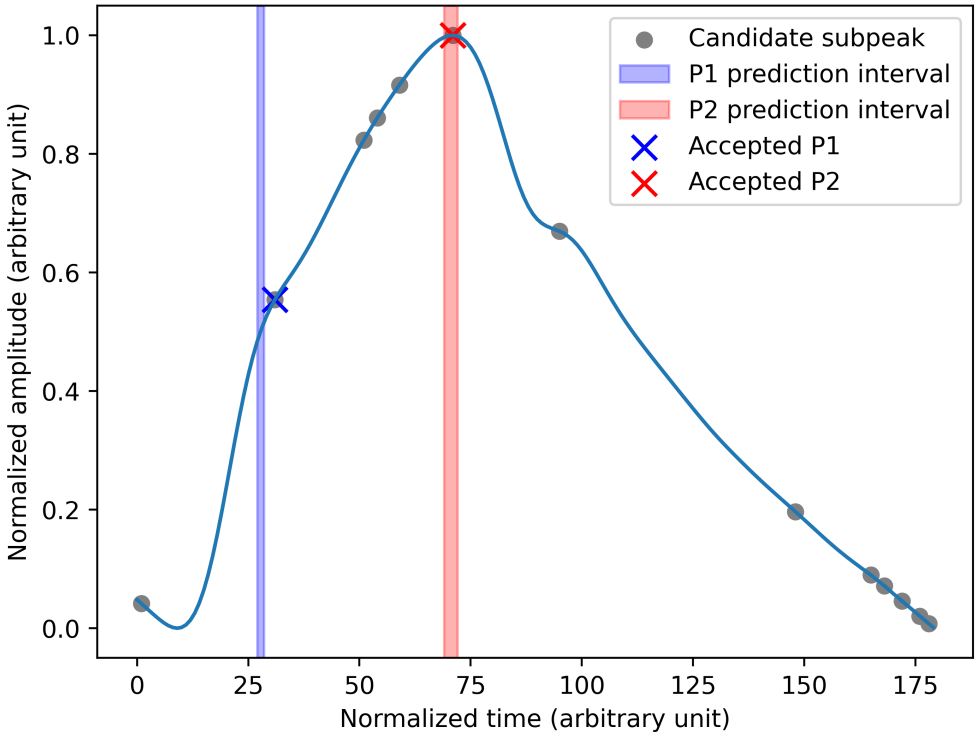

Figure 4: P1 and P2 designation process. The candidate subpeaks correspond to the positive local maxima of the curvature function. P1 and P2 prediction intervals are calculated by the dedicated neural network model. As described in this example, the chosen candidate may fall outside the prediction interval. The detected P1 and P2 are identical to those annotated by the expert.

2.6. Conformal Risk Control

2.6.1. Risk Control Procedure

In this step, we seek to guarantee several statistical properties of the pipeline's performance. The risk control procedure is designed to identify unreliable predictions in order to remove them from the final output. Three criteria are used to quantify the uncertainty of the neural network. Firstly, the classifier outputs the probability $\widehat{f}$ of having no P2/P1 ratio to compute (i.e., due to missing subpeaks or the presence of artifacts). Secondly, the subpeak detector outputs two research intervals, i.e. $[\ell_{p1}(x), u_{p1}(x)]$ and $[\ell_{p1}(x), u_{p1}(x)]$. In the following, their respective widths are denoted as $\widehat{w}_1$ and $\widehat{w}_2$. Consequently, thresholds on $\widehat{f}, \widehat{w}_1$ and $\widehat{w}_2$ can be used to reject predictions that are too uncertain and achieve statistical guarantees on the final pipeline performance. More formally, let $\lambda = [\lambda_0, \lambda_1, \lambda_2] \in [0,1] \times [0,180] \times [0,180]$ be a triplet of parameters. The final decision function $\mathcal{F}_\lambda$ is defined as follows:

$$\mathcal{F}_\lambda(x) = \begin{cases} -1, & \text{if } \widehat{f}(x) > \lambda_0 \text{ or } \widehat{w_1}(x) > \lambda_1 \text{ or } \widehat{w_2}(x) > \lambda_2 \\ \widehat{P}2(x)/\widehat{P}1(x), & \text{otherwise.} \end{cases}$$

We seek to simultaneously control both of the following risks:

$$R_0(\lambda) = \mathbb{P}(\mathcal{F}_\lambda(X_i) = -1 \,|\, Y_i \neq -1)$$

and

$$R_1(\lambda) = \mathbb{P}\left(\| \frac{Y_i - \mathcal{F}_\lambda(X_i)}{Y_i} \| > \epsilon \,|\, \mathcal{F}_\lambda(X_i) \neq -1 \right),$$

where $\| \cdot \|$ is the Euclidean norm. The probability R_0 corresponds to a valid pulse being rejected at any step of the detection pipeline. R_1 corresponds to the probability that the relative error on a displayed P2/P1 ratio will exceed a user-defined tolerance threshold ϵ, set to 10% in the present experiments. A P2/P1 ratio value displayed for a pulse without recognizable P1 and/or P2 (i.e., not belonging to the valid pulse class) is automatically considered as a failure. We seek to ensure that $\mathbb{P}(R_0 < \alpha_0) > 1 - \delta$ and that $\mathbb{P}(R_1 < \alpha_1) > 1 - \delta$. To do so, we look for a triplet λ that allows the rejection of both following null hypotheses:

$$\mathcal{H}_\lambda^0 : R_0(\lambda) > \alpha_0 \text{ and } \mathcal{H}_\lambda^1 : R_1(\lambda) > \alpha_1.$$

Under $\mathcal{H}_\lambda^0$ and $\mathcal{H}_\lambda^1$, R_0 and R_1 can be estimated using a calibration set of n pulses including the subset M of the pulses with a calculable ratio (i.e. with $Y \neq -1$):

$$\widehat{R}_{0,\text{null}}(\lambda) = \frac{1}{|M|} \sum_{X \in M} \mathbb{1}(\mathcal{F}_\lambda(X) = -1)$$

and

$$\widehat{R}_{1,\text{null}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{1}\left(\| \frac{Y_i - \mathcal{F}_\lambda(X_i)}{Y_i} \| > \epsilon \right) - \alpha_1 \mathbb{1}(\mathcal{F}_\lambda(X_i) = -1) \right],$$

where $|M|$ stands for the cardinal of the subset M and $\mathbb{1}(\cdot)$ denotes the indicator function.

Having defined a confidence level δ, we can compute the p-values p_λ^0 and p_λ^1, associated with $\mathcal{H}_\lambda^0$ and $\mathcal{H}_\lambda^1$, respectively. A controlling-risk triplet λ satisfies the criterion $p_\lambda = \max(p_\lambda^0, p_\lambda^1) < \delta$. Regarding the p-value computation, R_0 can be modeled as the mean of a sequence of iid Bernoulli variables. Consequently, p_λ^0 can be computed on the calibration set such that $p_\lambda^0 = B(\widehat{R}_{0,\text{null}}(\lambda))$, where B denotes the cumulative distribution function (CDF) of a binomial distribution $\mathcal{B}(|M|, \alpha_0)$. In contrast, p_λ^1 has to be calculated using the more conservative Hoeffding-Bentkus inequality p-value [24] due to its conditional component.

2.6.2. Search for Valid Thresholds

In practice, we limit our triplet λ research to the Cartesian product $\Lambda = [0.01, 0.02, ..., 0.99] \times [0.1, 0.2, ..., 3]^2$, leaving more than 80,000 null hypotheses to be tested at a confidence level $1 - \delta$. To alleviate the risk of rejecting a null hypothesis by error, the Learn-then-Test framework applies a family-wise error rate-(FWER) controlling algorithm to achieve the multiple hypothesis test. We opted for a fixed-sequence test procedure [25], which consists of sequentially testing a series of null hypotheses until one of them is accepted, i.e. its associated p-value is superior to δ. Afterwards, the set Λ^* of the thresholds associated with a rejected null hypothesis is considered as valid to achieve the desired risk control. The fixed-sequence test procedure implies that the most unlikely null hypotheses appear first in the test sequence. As in the case of three-dimensional thresholds, defining such a test sequence is not obvious, we relied on the split-sequence test procedure [16] to identify a hypothesis sequence to test. In a nutshell, this method involves randomly splitting the calibration dataset into two equal parts, identifying some of the most promising triplets in the first part, and then evaluating a testing path that prioritizes them in the second part

2.7. ***Baseline Comparison***

We built several variants of the analysis pipeline described above. In each of these, the LSTM-based networks were replaced by other machine learning algorithms. As with the original pipeline, the classification task was taught to a first model, while a second was used to perform a quantile regression. The chosen algorithms are the following:

- K-Neighbors (KNN): Let p be a pulse to analyze. For both classification and regression, the output of KNN is estimated by interpolating the labels of the k-nearest neighbors of p in the training set. Its extension to quantile regression can be done easily without a dedicated training step.
- Extreme Gradient Boosting (XGB): This algorithm is a L2-regularized version of the gradient boosted trees [26]. During the training step, a linear combination of decision trees is built iteratively to minimize a loss function. As any gradient descent-based algorithm, a quantile regression can be performed by using a pinball loss for the training step.
- Random Forests (RF): Its decision function relies on the combination of a big number of decision trees fitted on features randomly chosen during the training set. As RF do not rely on a gradient descent, the modifications necessary to perform a quantile regression are made at the evaluation step [26].
- Multi-Layer Perceptron (MLP): It corresponds to a simple neural network model where several fully-connected layer are stacked linearly. A 10%-dropout was applied before the output layer. As for the LSTM-based networks, the quantile regression was performed using a pinball loss during the training step.

For each algorithm, the main parameters were chosen using a 5-fold cross-validation on the training cohort as described in Table 2. Non-mentioned parameters were left by default.

Algorithm	Classification	Regression	Implementation
KNN	neighbors: 10, distance-weighted: yes	neighbors: 5, distance-weighted: no	Scikit-learn, sklearn-quantiles
RF	trees: 50	trees: 50	Scikit-learn, sklearn-quantiles
XGB	max depth: 6, L2-reg: 1.0	maximum depth: 6, L2-reg: 0.0	XGBoost
MLP	layer width: 16, hidden layers: 7	layer width: 16, hidden layers: 3	PyTorch

Table 2: Implementation details of the machine learning models used in the analysis pipeline

The detailed experiments are available at the following address: https://github.com/donatien-lege/ICP-SWAn-baseline

3. Results

3.1. Definition of the Rejection Thresholds

3.1.1. Raw Performance in the Validation Dataset

Raw metrics were first calculated on both parts of the pipeline (i.e., the pulse classification part and the subpeak detection part) using the validation dataset. The goal was to choose appropriate risk levels α_0 and α_1 for the calibration step. The results obtained with the classifier on the validation dataset (n = 2,862 pulses) are presented as a binary classification task. As the goal was to assess the ability of detecting pulses with a calculable P2/P1 ratio against the rest, classes 1 (non-interpretable pulses) and 2 (artifacts) have been merged. The receiver-operating characteristic (ROC) curve is presented in Figure 5A. Regarding the subpeak detector performance, the distribution of the relative errors in the P2/P1 ratio is presented in Figure 5B.

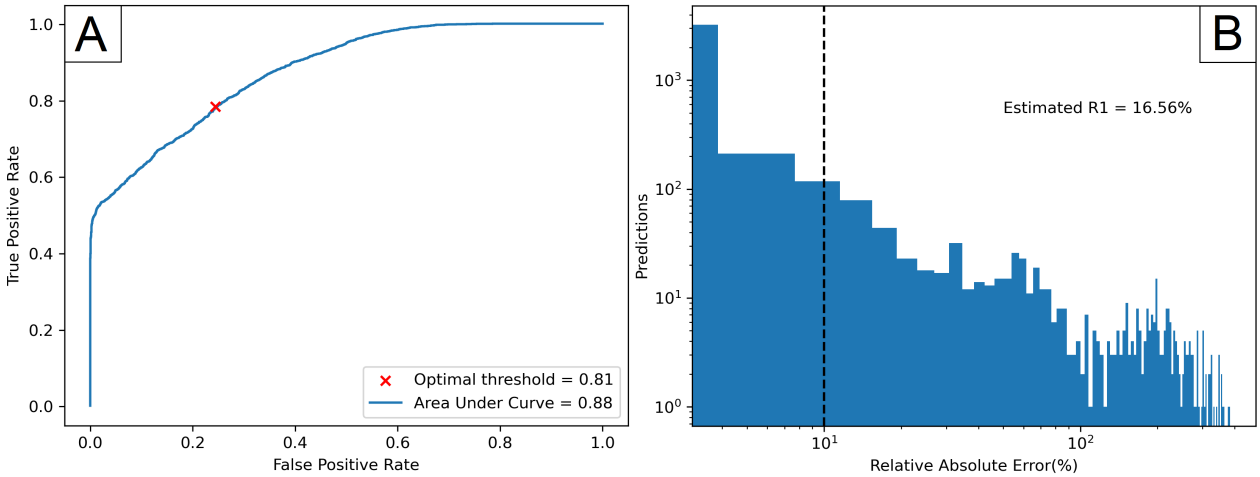

Figure 5: Raw performance of the P2/P1 calculation pipeline on the validation dataset. A: Receiver-Operating Characteristic Curve of the pulse selector. Metrics are calculated for a binary classification task where the positive class corresponds to the pulses with a calculable P2/P1 ratio and the negative class to non-interpretable pulses and artifacts combined. AUC — area under the curve. B: Distribution of the mean relative errors of the P2/P1 ratio detected by the subpeak designator.

A first naive threshold for the classifier can be chosen by maximizing the difference between the true positive rate and the false positive rate on the validation dataset. Using the ROC curve shown above, such a threshold was set to 0.81. Without additional criteria on w_1 and w_2, the corresponding triplet was defined as $\lambda^{init} = (0.81, 180, 180)$. $R_0(\lambda^{init})$ and $R_1(\lambda^{init})$ were estimated as 1.3% and 16.6%, respectively.

3.1.2. Threshold Calculation

For a given data distribution, it is in general not possible to control any risk at any level α. Therefore, the performance obtained in the validation dataset is used as an aid to choose realistic levels of risk control. We chose our tolerance levels by rounding up to the nearest 5% the $\widehat{R}_0$ and $\widehat{R}_1$ estimated with the validation dataset. Hence, the tolerance levels α_0 and α_1 were set to 5% and 15%, respectively. A further calibration was also performed for less restrictive thresholds, i.e. $\alpha_0 = 10\%$ and $\alpha_1 = 20\%$. The aim was to ensure that risk control could be achieved for each model, but also to gain insight into the impact of the choice of threshold.

For each calibration, the confidence level δ was set to 0.05. The calibration dataset was split into two equal parts C_1 and C_2. A sequence of 200 hypotheses was built from C_1 to perform a fixed-sequence testing procedure performed on C_2. Among the threshold triplets that allowed the rejection of both null hypotheses, we chose the triplet that appeared last before the null hypotheses were accepted.

3.2. Model Comparison

Several performance indicators are compared between the two test datasets. In addition to the estimated risks R_0 and R_1 used in the calibration procedure, (i) the mean absolute error of the P2/P1 ratio ($MAE_{P2/P1}$), (ii) the proportion of pulses where the absolute delay between the actual P1 and the predicted P1 exceeded 10ms ($\Delta P1_{>10\text{ ms}}$), and (iii) the proportion of pulses where the absolute delay between the actual P2 and the predicted P2 exceeded 10ms ($\Delta P2_{>10\text{ ms}}$). To assess the impact of the data augmentation step, we also trained a full LSTM-based pipeline, denoted LSTM*, without data augmentation.

3.2.1. Public Test Dataset

As seen in Table 3 the naive $\hat{R}_0$ and $\hat{R}_1$ are already far below the user-defined tolerance levels for most models. The calibration step was slightly detrimental for the neural-network based models (i.e., MLP and LSTM) by increasing the $\widehat{R_0}$ without real improvements on the other metrics. In contrast, it helped stabilize the decision rules of the tree-based models by lowering $\widehat{R_0}$ below 1%. Naive thresholds chosen for LSTM* and KNN were clearly too restrictive with $\widehat{R_0}$ exceeding 15%. Interestingly, the two LSTM-based pipelines achieved the subpeak detection with the best precision on P1 ($\Delta P1_{>10\text{ ms}} < 5\%$) and the worst precision on P2 ($\Delta P2_{>10\text{ ms}} > 14\%$).

Further analysis can be done by grouping the metrics by pulse shape to identify the main error sources. The detailed errors are presented in Table 4 to compare the performances of the LSTM-based network before and after the calibration to $\alpha_0 = 5\%$ and $\alpha_1 = 15\%$. The calibration step mostly impacts the pipeline's outputs for the pulses of shapes "T1" and "A+E" for which the false rejects are more frequent. In particular, the $\widehat{R_0}$ for the class "A+E" goes from 8.17% to 17.98%.

3.2.2. Custom Test Dataset

In order to validate the inter-annotator agreement, 4 ICP recordings accounting for 514 pulses were independently labeled by three annotators (VG, GP, AK). The classification was unanimous in 94.9% of the cases. For 481 pulses (91.9%), at least two out of three P2/P1 ratios were calculated. The average absolute difference between the P2/P1 ratios was 0.02 (std = 0.07). The results are presented in Table 5.

In this cohort, it is clear that the risk control could not be achieved for $\alpha_0 = 5\%$ and $\alpha_1 = 15\%$. Indeed, with the exception of the LSTM, all the pipelines exceeded the thresholds chosen for R_0 and R_1. However, the calibration procedure helped homogenize the performances across the different pipelines, especially in the cases of KNN and RF for which the naive ROC-based thresholds were associated with high false negative rate (55.31% and 19.37%, respectively). As for the public test dataset, the different metrics grouped by pulse shape are presented in Table 6.

In this more challenging cohort, the calibration step still had a cost on the false rejects (the R_0 passed from 1.68% to 4.94%) but also improved all the other metrics. As in the public dataset, the shape "A+E" is the most impacted with a R_0 going from 23.11% to 34.75%).

	Calibration	$\widehat{R_0}$ (%)	$\widehat{R_1}$ (%)	$\Delta P1_{>10\text{ ms}}$ (%)	$\Delta P2_{>10\text{ ms}}$ (%)	$MAE_{P2/P1}$
KNN	C_{ROC}	39.20	1.86	5.77	6.92	0.015
	C_{5-15}	0.32	3.24	9.62	13.05	0.024
	C_{10-20}	1.92	2.98	9.36	12.59	0.023
RF	C_{ROC}	6.15	2.99	8.92	11.11	0.02
	C_{5-15}	0.72	3.62	9.76	12.40	0.023
	C_{10-20}	2.46	2.85	9.51	11.53	0.019
XGB	C_{ROC}	2.66	3.01	9.44	11.61	0.02
	C_{5-15}	0.21	3.44	9.88	12.27	0.022
	C_{10-20}	0.91	3.23	9.79	11.98	0.021
MLP	C_{ROC}	0.97	2.89	9.40	12.94	0.02
	C_{5-15}	8.41	2.93	8.85	12.04	0.020
	C_{10-20}	0.30	2.99	9.57	13.06	0.021
LSTM	C_{ROC}	0.36	1.89	3.70	14.94	0.016
	C_{5-15}	3.52	1.88	3.71	14.70	0.016
	C_{10-20}	7.38	1.88	3.71	14.43	0.016
LSTM*	C_{ROC}	17.44	4.86	4.14	18.90	0.043
	C_{5-15}	7.11	4.92	4.64	18.23	0.043
	C_{10-20}	9.83	5.06	4.72	18.58	0.044

Table 3: Model comparison on the public test dataset. $MAE_{P2/P1}$ — Mean Absolute Error on the calculated P2/P1 ratio. $\widehat{R_0}$ — proportion of the pulses with a calculable P2/P1 ratio rejected by error. $\widehat{R_1}$ — fraction of the accepted pulses without a P2/P1 ratio or whose mean relative error on the P2/P1 ratio is greater than 10%. $\Delta P1_{>10\text{ ms}}$ — Proportion of the valid pulses whose P1 detection delay exceeds 10 ms. $\Delta P2_{>10\text{ ms}}$ — Proportion of the valid pulses whose P2 detection delay exceeds 10 ms. KNN — K-Nearest Neighbors, RF — Random Forests, XGB — eXtreme Gradient Boosting, MLP — MultiLayer Perceptron, LSTM — Long Short-Term Memory, LSTM* — LSTM-based detection pipeline trained without data augmentation. C_{ROC} — Naive calibration. C_{5-15} — Calibration performed with $\alpha_0 = 5\%$ and $\alpha_1 = 15\%$. C_{10-20} — Calibration performed with $\alpha_0 = 10\%$ and $\alpha_1 = 20\%$.

	Naive thresholds (C_{ROC})					Calibrated thresholds (C_{5-15})				
Shape	$\widehat{R_0}$(%)	$\widehat{R_1}$(%)	$\Delta P1_{>10\text{ ms}}$	$\Delta P2_{>10\text{ ms}}$	$\text{MAE}_{\text{P2/P1}}$	$\widehat{R_0}$(%)	$\widehat{R_1}$(%)	$\Delta P1_{>10\text{ ms}}$	$\Delta P2_{>10\text{ ms}}$	$\text{MAE}_{\text{P2/P1}}$
T1	0.34	2.36	0.41	14.11	0.010	9.96	2.24	0.37	15.17	0.010
T2	0.03	0.60	3.46	13.84	0.008	2.65	0.58	3.43	13.33	0.008
T3	0.11	0.71	1.99	10.76	0.014	0.71	0.72	2.01	10.46	0.013
T4	0.0	18.09	21.98	34.05	0.013	0.00	18.09	21.98	34.05	0.134
A+E	8.17	7.72	8.90	43.92	0.029	17.98	8.64	8.31	45.18	0.029
all	0.36	1.89	3.70	14.94	0.016	3.52	1.88	3.71	14.70	0.016

Table 4: Comparison of the performance metrics of the LSTM-based P2/P1 calculation pipeline using the naive and calibrated thresholds in the public test dataset. $MAE_{P2/P1}$ — Mean Absolute Error on the calculated P2/P1 ratio. $\widehat{R_0}$ — proportion of the pulses with a calculable P2/P1 ratio rejected by error. $\widehat{R_1}$ — fraction of the accepted pulses without a P2/P1 ratio or whose mean relative error on the P2/P1 ratio is greater than 10%. $\Delta P1_{>10\text{ ms}}$ — Proportion of the valid pulses whose P1 detection delay exceeds 10 ms. $\Delta P2_{>10\text{ ms}}$ — Proportion of the valid pulses whose P2 detection delay exceeds 10 ms. C_{ROC} — Naive calibration. C_{5-15} — Calibration performed with $\alpha_0 = 5\%$ and $\alpha_1 = 15\%$.

In the other pulse shapes, the calibration had a homogeneous effect, increasing the accuracy of subpeak detection at the cost of a slightly higher rate of erroneous rejects.

	Calibration	$\widehat{R_0}$ (%)	$\widehat{R_1}$ (%)	$\Delta P1_{>10\text{ ms}}$ (%)	$\Delta P2_{>10\text{ ms}}$ (%)	$MAE_{P2/P1}$
KNN	C_{ROC}	55.31	9.18	17.43	21.38	0.042
	C_{5-15}	0.93	19.89	20.44	26.91	0.087
	C_{10-20}	3.90	18.59	20.22	26.28	0.085
RF	C_{ROC}	19.37	14.08	19.72	27.960	0.061
	C_{5-15}	1.62	20.51	21.01	30.24	0.085
	C_{10-20}	4.44	18.83	20.33	29.04	0.079
XGB	C_{ROC}	5.57	19.21	21.81	30.64	0.094
	C_{5-15}	0.38	22.33	22.50	31.52	0.101
	C_{10-20}	1.93	21.26	22.20	31.06	0.100
MLP	C_{ROC}	6.28	17.02	20.48	25.61	0.076
	C_{5-15}	10.73	15.38	20.12	24.49	0.067
	C_{10-20}	0.90	20.43	21.09	26.97	0.088
LSTM	C_{ROC}	1.68	14.51	7.32	14.39	0.063
	C_{5-15}	4.94	13.52	6.74	13.87	0.061
	C_{10-20}	10.68	12.85	6.60	12.900	0.060
LSTM*	C_{ROC}	7.94	17.23	9.49	15.09	0.081
	C_{5-15}	4.67	17.96	9.77	15.53	0.09
	C_{10-20}	5.37	17.73	9.59	15.35	0.088

Table 5: Comparison of the performance metrics of the P2/P1 calculation pipeline using the naive and calibrated thresholds in the public test dataset. $MAE_{P2/P1}$ — Mean Absolute Error on the calculated P2/P1 ratio. $\widehat{R_0}$ — proportion of the pulses with a calculable P2/P1 ratio rejected by error. $\widehat{R_1}$ — fraction of the accepted pulses without a P2/P1 ratio or whose mean relative error on the P2/P1 ratio is greater than 10%. $\Delta P1_{>10\text{ ms}}$ — Proportion of the valid pulses whose P1 detection delay exceeds 10 ms. $\Delta P2_{>10\text{ ms}}$ — Proportion of the valid pulses whose P2 detection delay exceeds 10 ms. C_{ROC} — Naive calibration. C_{5-15} — Calibration performed with $\alpha_0 = 5\%$ and $\alpha_1 = 15\%$.

	Naive thresholds (C_{ROC})					Calibrated thresholds (C_{5-15})				
Shape	$\widehat{R_0}$(%)	$\widehat{R_1}$(%)	$\Delta P1_{>10\text{ ms}}$	$\Delta P2_{>10\text{ ms}}$	$\text{MAE}_{\text{P2/P1}}$	$\widehat{R_0}$(%)	$\widehat{R_1}$(%)	$\Delta P1_{>10\text{ ms}}$	$\Delta P2_{>10\text{ ms}}$	$\text{MAE}_{\text{P2/P1}}$
T1	0.54	18.48	2.42	28.17	0.037	4.76	18.16	2.08	28.02	0.034
T2	0.71	9.42	7.29	13.93	0.031	4.51	8.74	6.54	13.22	0.028
T3	0.43	3.98	4.04	5.77	0.028	2.25	3.16	3.60	5.61	0.026
T4	0.54	35.20	20.16	18.44	0.25	2.49	34.96	20.16	18.17	0.253
A+E	23.11	43.04	23.32	30.77	0.18	34.75	40.64	20.96	28.89	0.167
all	1.68	14.51	7.32	14.39	0.063	4.94	13.52	6.74	13.87	0.061

Table 6: Comparison of the performance metrics of the LSTM-based P2/P1 calculation pipeline using the naive and calibrated thresholds in the custom test dataset. $MAE_{P2/P1}$ — Mean Absolute Error on the calculated P2/P1 ratio. $\widehat{R_0}$ — proportion of the pulses with a calculable P2/P1 ratio rejected by error. $\widehat{R_1}$ — fraction of the accepted pulses without a P2/P1 ratio or whose mean relative error on the P2/P1 ratio is greater than 10%. $\Delta P1_{>10\text{ ms}}$ — Proportion of the valid pulses whose P1 detection delay exceeds 10 ms. $\Delta P2_{>10\text{ ms}}$ — Proportion of the valid pulses whose P2 detection delay exceeds 10 ms. C_{ROC} — Naive calibration. C_{5-15} — Calibration performed with $\alpha_0 = 5\%$ and $\alpha_1 = 15\%$.

3.3. Continuous P2/P1 evaluation

To produce real-world examples of continuous P2/P1 calculation with the trained models, ICP-SWAn was retrospectively run over the two first days of ICP monitoring in the testing cohort. The training cohort was left aside for two reasons: (i) about a half of the recordings only lasted a few hours, (ii) the proportion of missing values could be underestimated on signals used to train the models. In parallel to the P2/P1 ratio, we also calculated the pulse

amplitude (AMP) and the pulse shape index (PSI) based on the automated morphological classification. The PSI correspond to average class of the cardiac pulses included in a 5-min average window with a 10-second shift. For sake of simplicity, we applied the same moving average to the P2/P1 ratio and to the amplitude. The mean value of the window was calculated only if it contained less than 50% of missing values. An example of such a retrospective monitoring is given in Figure 6.

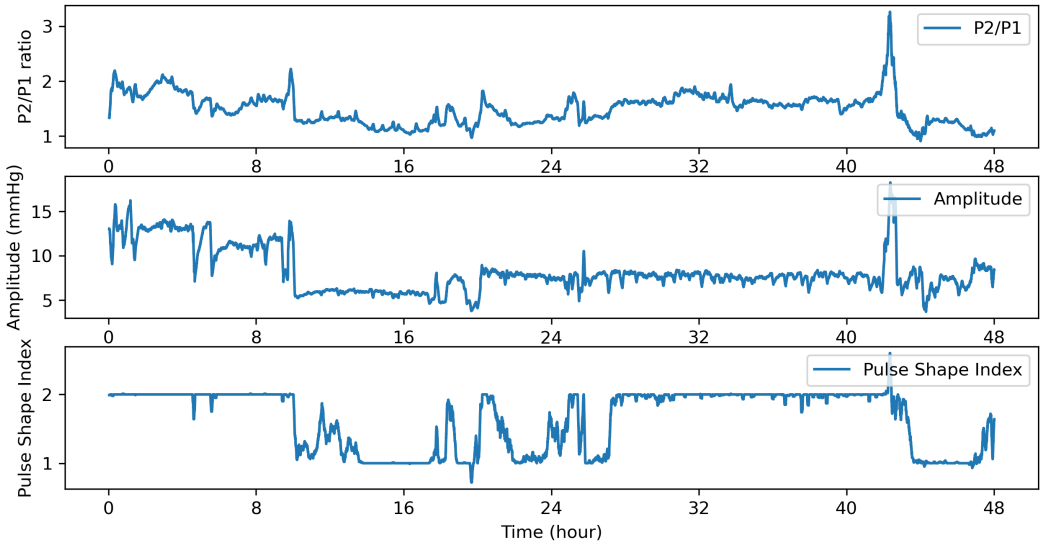

Figure 6: Example of a 2-day retrospective intracranial compliance monitoring

In practice, the P2/P1 monitoring can be discontinuous if too many pulses are rejected. As shown in Figure 7 A, the amount of time without P2/P1 can vary a lot between different patients. Among the 49 investigated monitorings, the proportion of missing P2/P1 values exceeded 10% in a dozen of cases. Only one highly noisy monitoring contained more than 50% of missing values. Interestingly, AMP was weakly correlated with PSI (0.15) and P2/P1 ratio (0.22) (see Figure 7 B).

Regarding the feasibility in a real-time setting, an on-boarded version of ICP-SWAn was implemented on a 600-MHz i.MX6 microprocessor without a graphics card. The ICP signal was analyzed by 1-min ICP batches. The average time of analyzing a batch was 2.32 sec (max: 5.5 sec). In comparison, on a regular laptop equipped with a GPU (The exact specifications are given in Section 2.5), the average time for processing a batch was 0.028 sec (max: 0.040 sec).

4. Discussion

With ICP-SWAn, we proposed a P1 and P2 subpeak detection approach for the ICP signal that incorporates selective regression to assess the confidence of predictions and exclude uncertain cases from the final output. The main goal is to avoid displaying results that are likely to be erroneous in an environment that is already saturated with information, such as ICUs. In practice, our ICP waveform analysis pipeline is designed to isolate individual pulses from the input univariate ICP signal and return the associated P2/P1 ratio values

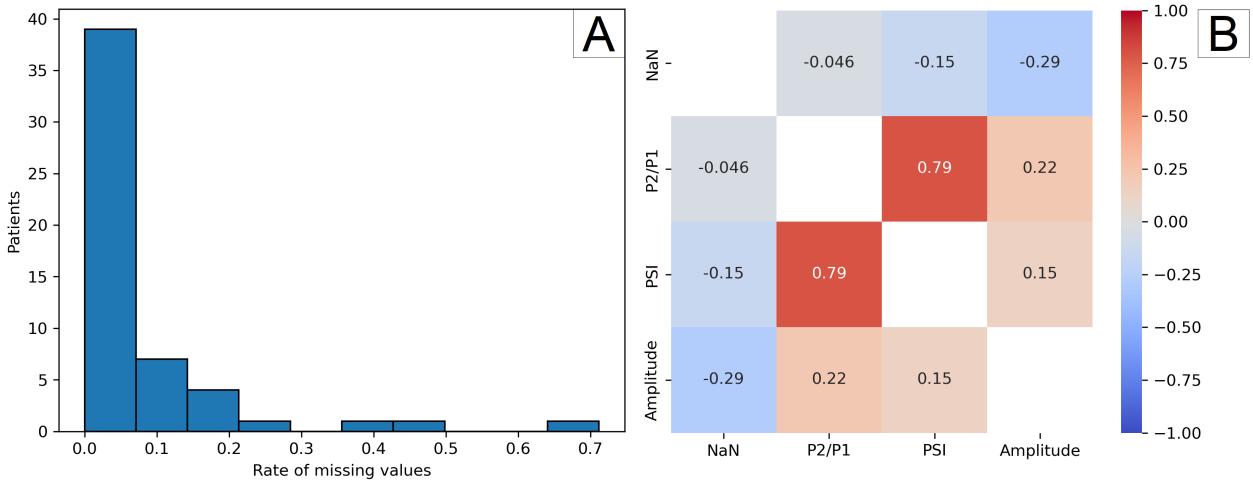

Figure 7: A: Proportion of missing P2/P1 values among 49 2-day monitorings. B: Correlation map between the proportion of missing values (NaN) and three indicators of intracranial compliance.

when the confidence is high enough. To do this, a calibration procedure was performed on a dedicated pulse subset to simultaneously control two risks at a confidence level of 95%. The expectations for both risks were determined based on the performances of the non-calibrated pipeline on a validation subset. Firstly, we wanted to guarantee that the probability $\widehat{R}_0$ of rejecting a pulse with a calculable P2/P1 ratio wass less than 5%. Secondly, we wanted to guarantee that the probability $\widehat{R}_1$ of returning a highly erroneous P2/P1 ratio (i.e., associated with a relative error $> 10\%$ or calculated on an invalid pulse) was less than 15%. As the calibration procedure is model-agnostic, we also compared our LSTM-based pipeline with four other machine learning (ML) algorithms able to perform a quantile regression. The quantile estimation can either be performed directly at the evaluation phase (such as for KNN or RF) or by using an asymmetric loss during the training phase (XGB, LSTM and MLP). Since the risk control cannot be achieved for any ML algorithm at any threshold, we also compared the performances of the algorithm on a less restrictive calibration to ensure that the full calibration procedure could be run adequately on each ML algorithm (seeking to guarantee that $\widehat{R}_0 < 10$ % and $\widehat{R}_1 < 20$ %). Performances were evaluated on two distinct datasets, comprised of a total of 113 patients admitted to the NICU. The first dataset was a 11,172-pulse public dataset, whereas the second one, consisting of 13,086 pulses, was selected to represent an important proportion of the difficulties encountered in real-time monitoring scenarios and included 10.8% of artifacts and 6.4% of pulses without a calculable P2/P1 ratio. A reference AI-based morphological classification [18] was used to assess the waveform diversity in both datasets. As presented in Table 1, the public dataset exhibited comparatively lower ICP levels (3.46 mmHg vs 8.88 mmHg, on average), lower P2/P1 ratios (1.25 vs 1.44, on average), and a higher proportion of pulses of type T2 - "possibly pathological" with clearly visible P1 but elevated P2 (53.7% vs 24.7%). We considered that the influence of the sampling rate in the acquisition of the different ICP signals were negligible on the pipeline performances for two reasons: (i) all the signals were acquired with a sampling rate superior to 50Hz [27] and (ii) the pulses were interpolated to 180 points at the preprocessing step.

On the public test dataset, the calibration step tended to homogenize the performances of all the pipelines, although neural network-based ones exhibited slightly higher false negative

rates ($R_0 = 3.52\%$). The MAE on the P2/P1 comprised between 0.016 (LSTM) and 0.023 (KNN and RF). In the specific case of the LSTM-based pipeline, the calibration procedure had little to no effect on detection performances, but increased by 3.12% the percentage of excluded pulses. The results can be explained by the baseline values of $\widehat{R}_0$ and $\widehat{R}_1$ in this dataset that were already far below the defined tolerance thresholds. If all of the detection pipelines met the statistical expectations, it is noticeable that the data augmentation step exhibited clear benefits on the performances of the calibrated LSTM-based pipeline, with R_0 passing from 7.11% to 3.52% and R_1 passing from 4.92% to 1.88%.

In contrast, our custom test dataset appeared to be more difficult to process. In this setup, $\widehat{R}_0$ and $\widehat{R}_1$ were estimated as 4.92% and 13.52%, respectively, which appeared to be the best compromise between $\widehat{R}_0$ and $\widehat{R}_1$ across all the trained pipelines. Retrospectively, the expected risk control was only achieved for the LSTM pipeline. Relaxing the two thresholds to 10% and 20%, respectively, allowed us to run the calibration procedure adequately on each pipeline, by building an appropriate testing path with contrasted p-values among the tested triplets. However, for all the ML algorithms, it appeared that maintaining the $\widehat{R}_1$ below 15% could only be achieved with a significant cost on the negative rate. Overall, errors were mainly concentrated in the pulse shape classes T1 ("normal"), T4 ("likely pathological") and A+E ("artifacts or errors"). T1 corresponds to the pulses associated with the highest ICC, where P1 is dominant and P2 is often less pronounced. In contrast, T4 corresponds to triangular waveforms in which the exact position of P1 can be difficult to locate. Finally, pulses of type A+E are close to the rejection boundary, and are therefore often difficult to interpret. As a result, for both data sets, the calibration process first increased the proportion of false rejects in the A+E class. In contrast, all the other metrics improved evenly across the remaining classes. In this sense, calibrating mainly helped to identify a few outlying prediction. In addition, false rejections mostly occur when the human decision itself is uncertain.

A major difficulty in this study was presented by the annotation process. Marking subpeaks in the ICP signal may seem easy enough and largely independent of the operator, especially when synchronized with the ABP signal that facilitates the positioning of P1. Unfortunately, some cases remain difficult to interpret and there is no way to assess the underlying physiological mechanisms; even though all the annotators have reported the difficulty in labeling some of the ICP segments (10 out of 98 were even re-labeled independently by all the annotators), the overall level of human confidence is not easy to quantify. In practice, these difficult cases are not distributed evenly among the ICP recordings. Since changes in the ICP waveform morphology are generally slow and progressive (at the scale of hours), uninterpretable waveforms tend to occur in a limited subset of patients, but for a significant period of time, compromising the use of this information in these situations.

Learn-then-Test methods were added to the pipeline to identify unreliable predictions. In other words, we tried to find an optimal trade-off between the risk of excessive abstention from subpeak detection and the risk of misleading P2/P1 results. In the context of intensive care, we believe that a decision that leans toward the former risk is more acceptable. In addition, as the ICP waveform is generally averaged over several minutes [28], isolated erroneous pulse deletions have little to no effect on the information given to the clinician. Throughout our P2/P1 ratio calculation pipeline, three quality checks are performed to ensure the validity of a prediction. The first one corresponds to the classifier output, whereas the second and the third criteria rely on the widths of the respective prediction intervals for P1 and P2. In the future, more complex criteria can be defined and integrated into the calibration step in the same way. Furthermore, it would be possible to define additional risks to be controlled.

We chose to focus on a proportion of misleading values (i.e., those with an elevated relative P2/P1 error or corresponding to non-identifiable P1 and P2) to account for the occurrence of artifacts, but other criteria could be chosen, for example, based on the MAE of the P2/P1 ratio. The decision to control a relative error is driven by the clinical interpretation of the P2/P1 ratio: more tolerance can be shown with higher values since the waveform is characteristic of an impaired ICC anyway. Overall, calibration becomes more interesting when dealing with results close to predefined statistical limits. Due to its effect on the false-negative rate, its use is mainly suitable in a real-time setup where out-of-distribution cases may occur without the possibility of human correction.

Our experiments on continuous ICP monitorings show that purely morphological indices such as PSI and P2/P1 ratio only weakly correlate with AMP. Over 49 2-day recordings, the correlation did not exceed 0.22 between the mean P2/P1 ratio and the mean AMP, which confirm that these indices do not exactly reflect the same biomechanical properties of the brain [29]. In contrast, mean PSI and mean P2/P1 ratio showed a good agreement with a Pearson correlation of 0.79. However, as shown in Figure 6, the P2/P1 ratio can show more "physiological" variations along the monitoring due to its continuous nature. Further investigations are needed to assess the clinical relevance of these subtle modifications in the ICP waveform.

5. Conclusion

We have developed the ICP-SWAn pipeline to calculate the P2/P1 ratio on univariate ICP signals in real time. Three quality checks are performed on each prediction to identify uncertain situations where no result should be displayed to the clinician. The algorithm's performance was assessed both on a 64-patient public benchmark dataset and on a 49-patient custom dataset, where artifacts and pulses with missing peaks were over-represented compared to raw ICP recordings. Errors occurred primarily in ICP waveforms associated with both ends of the intracranial compliance spectrum. Overall performance was consistent with the *a priori* statistical expectations.

References

[1] M. C. Cnossen, J. A. Huijben, M. van der Jagt, V. Volovici, T. van Essen, S. Polinder, D. Nelson, A. Ercole, N. Stocchetti, G. Citerio, W. C. Peul, A. I. R. Maas, D. Menon, E. W. Steyerberg, H. F. Lingsma, Variation in monitoring and treatment policies for intracranial hypertension in traumatic brain injury: a survey in 66 neurotrauma centers participating in the CENTER-TBI study, Critical Care 21 (1) (2017) 233. doi:10.1186/s13054-017-1816-9.
URL http://ccforum.biomedcentral.com/articles/10.1186/s13054-017-1816-9

[2] G. Cucciolini, V. Motroni, M. Czosnyka, Intracranial pressure for clinicians: It is not just a number, Journal of Anesthesia, Analgesia and Critical Care 3 (1) (2023) 31.

[3] M. Czosnyka, J. D. Pickard, Monitoring and interpretation of intracranial pressure, Journal of Neurology, Neurosurgery & Psychiatry 75 (6) (2004) 813–821.

[4] A. A. Domogo, P. Reinstrup, J. T. Ottesen, Mechanistic-mathematical modeling of intracranial pressure (icp) profiles over a single heart cycle. the fundament of the icp curve form, Journal of Theoretical Biology 564 (2023) 111451.

[5] E. Carrera, D.-J. Kim, G. Castellani, C. Zweifel, Z. Czosnyka, M. Kasprowicz, P. Smielewski, J. D. Pickard, M. Czosnyka, What shapes pulse amplitude of intracranial pressure?, Journal of Neurotrauma 27 (2) (2010) 317–324.

[6] J.-Y. Fan, C. Kirkness, P. Vicini, R. Burr, P. Mitchell, Intracranial pressure waveform morphology and intracranial adaptive capacity, American Journal of Critical Care 17 (6) (2008) 545–554.

[7] M. Czosnyka, Z. Czosnyka, Origin of intracranial pressure pulse waveform, Acta Neurochirurgica 162 (2020) 1815–1817.

[8] A. Kazimierska, M. Kasprowicz, M. Czosnyka, M. M. Placek, O. Baledent, P. Smielewski, Z. Czosnyka, Compliance of the cerebrospinal space: comparison of three methods, Acta Neurochirurgica 163 (2021) 1979–1989.

[9] A. Islam, L. Froese, T. Bergmann, A. Gomez, A. S. Sainbhi, N. Vakitbilir, K. Y. Stein, I. Marquez, Y. Ibrahim, F. A. Zeiler, Continuous monitoring methods of cerebral compliance and compensatory reserve: a scoping review of human literature, Physiological Measurement (2024).

[10] C. Zhang, S.-Y. Long, W.-d. You, G.-Y. Gao, X.-F. Yang, The value of the correlation coefficient between the icp wave amplitude and the mean icp level (rap) combined with the resistance to csf outflow (rout) for early prediction of the outcome before shunting in posttraumatic hydrocephalus, Frontiers in Neurology 13 (2022) 881568.

[11] M. Rozanek, J. Skola, L. Horakova, V. Trukhan, Effect of artifacts upon the pressure reactivity index, Scientific Reports 12 (1) (2022) 15131.

[12] P. Rashidinejad, X. Hu, S. Russell, Patient-adaptable intracranial pressure morphology analysis using a probabilistic model-based approach, Physiological Measurement 41 (10) (2020) 104003.

[13] X. Hu, P. Xu, F. Scalzo, P. Vespa, M. Bergsneider, Morphological clustering and analysis of continuous intracranial pressure, IEEE Transactions on Biomedical Engineering 56 (3) (2008) 696–705.

[14] H.-J. Lee, E.-J. Jeong, H. Kim, M. Czosnyka, D.-J. Kim, Morphological feature extraction from a continuous intracranial pressure pulse via a peak clustering algorithm, IEEE Transactions on Biomedical Engineering 63 (10) (2015) 2169–2176.

[15] K. Lewandowska, M. Weisbrot, A. Cieloszyk, W. Medrzycka-Dabrowska, S. Krupa, D. Ozga, Impact of alarm fatigue on the work of nurses in an intensive care environment—a systematic review, International Journal of Environmental Research and Public Health 17 (22) (2020) 8409.

[16] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, L. Lei, Learn then test: Calibrating predictive algorithms to achieve risk control, arXiv preprint arXiv:2110.01052 (2021).

[17] D. Legé, L. Gergelé, M. Prud'homme, J.-C. Lapayre, Y. Launey, J. Henriet, A deep learning-based automated framework for subpeak designation on intracranial pressure signals, Sensors 23 (18) (2023) 7834.

[18] C. Mataczyński, A. Kazimierska, A. Uryga, M. Burzyńska, A. Rusiecki, M. Kasprowicz, End-to-end automatic morphological classification of intracranial pressure pulse waveforms using deep learning, IEEE Journal of Biomedical and Health Informatics 26 (2) (2021) 494–504.

[19] N. Carney, A. M. Totten, C. O'Reilly, J. S. Ullman, G. W. Hawryluk, M. J. Bell, S. L. Bratton, R. Chesnut, O. A. Harris, N. Kissoon, A. M. Rubiano, L. Shutter, R. C. Tasker, M. S. Vavilala, J. Wilberger, D. W. Wright, J. Ghajar, Guidelines for the Management of Severe Traumatic Brain Injury, Fourth Edition, Neurosurgery 80 (1) (2017) 6–15. doi:10.1227/NEU.0000000000001432.
URL https://journals.lww.com/00006123-201701000-00003

[20] M. Wei, S. Krakauskaite, S. Subramanian, F. Scalzo, Peak detection in intracranial pressure signal waveforms: a comparative study, BioMedical Engineering OnLine 23 (1) (2024) 61.

[21] Z. Ahmed, F. Chaudhary, M. P. Fraix, D. K. Agrawal, Epidemiology, pathophysiology, and treatment strategies of concussions: a comprehensive review, Fortune journal of health sciences 7 (2) (2024) 197.

[22] B. Lv, J.-X. Lan, Y.-F. Si, Y.-F. Ren, M.-Y. Li, F.-F. Guo, G. Tang, Y. Bian, X.-H. Wang, R.-J. Zhang, et al., Epidemiological trends of subarachnoid hemorrhage at global, regional, and national level: a trend analysis study from 1990 to 2021, Military Medical Research 11 (1) (2024) 46.

[23] S. M. Bishop, A. Ercole, Multi-scale peak and trough detection optimised for periodic and quasi-periodic neuroscience data, in: Intracranial Pressure & Neuromonitoring XVI, Springer, 2018, pp. 189–195.

[24] S. Bates, A. Angelopoulos, L. Lei, J. Malik, M. Jordan, Distribution-free, risk-controlling prediction sets, Journal of the ACM (JACM) 68 (6) (2021) 1–34.

[25] B. L. Wiens, A fixed sequence bonferroni procedure for testing multiple endpoints, Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry 2 (3) (2003) 211–215.

[26] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al., Xgboost: extreme gradient boosting, R package version 0.4-2 1 (4) (2015) 1–4.

[27] S. Holm, P. K. Eide, Impact of sampling rate for time domain analysis of continuous intracranial pressure (icp) signals, Medical engineering & physics 31 (5) (2009) 601–606.

[28] A. Kazimierska, A. Uryga, C. Mataczyński, M. Czosnyka, E. W. Lang, M. Kasprowicz, Relationship between the shape of intracranial pressure pulse waveform and computed tomography characteristics in patients after traumatic brain injury, Critical Care 27 (1) (2023) 447.

[29] M. Kasprowicz, C. Mataczyński, A. Uryga, A. I. Pelah, E. Schmidt, M. Czosnyka, A. Kazimierska, Impact of age and mean intracranial pressure on the morphology of intracranial pressure waveform and its association with mortality in traumatic brain injury, Critical Care 29 (1) (2025) 1–12.