Functional Age Estimation Through Neonatal Motion Characterization Using Continuous Video Recordings

Sandie Cabon, Raphaël Weber, Antoine Simon, Patrick Pladys, Fabienne Porée, Guy Carrault

Functional age estimation through neonatal motion characterization using continuous video recordings

Sandie Cabon, Raphaël Weber, Antoine Simon, Patrick Pladys, Fabienne Porée and Guy Carrault

Univ Rennes, CHU Rennes, INSERM, LTSI - UMR 1099, F-35000 Rennes, France

Abstract: The follow-up of the development of the premature baby is a major component of its clinical care since it has been shown that it can reveal a pathology. However, no method allowing an automated and continuous monitoring of this development has been proposed. Within the framework of the Digi-NewB European project, our team wishes to offer new clinical indices qualifying the maturation of newborns. In this study, we propose a new method to characterize motor activity from video recordings. For this purpose, we have chosen to characterize the motion temporal organization by drawing inspiration from sleep organization. Thus, we propose a fully automatic process allowing to extract motion features and to combine them to estimate a functional age. By investigating two datasets, one of 28.5 hours (manually annotated) from 33 newborns and one of 4,920 hours from 46 newborns, we show that the proposed approach is relevant for monitoring in clinical routine and that the extracted features reflect the maturation of preterm newborns. Indeed, a compact and interpretable model using gestational age and three motion features (mean duration of intervals with motion, total percentage of time spent in motion and number of intervals without motion) was designed to predict post-menstrual age of newborns and showed an admissible mean absolute error of 1.3 weeks. While the temporal organization of motion was not studied clinically due to a lack of technological means, these results open the door to new developments, new investigations and new knowledge on the evolution of motion in newborns.

Keywords: video, motion, premature newborns, neonatal intensive care unit, monitoring, neurobehavioral development, machine learning.

1 Introduction

In recent years, the collection of large volumes of data associated with machine learning algorithms made possible to investigate new decision support tools. Today, there is a growing interest to bring these technologies into the clinic to improve patient care [1]. In this context, the European project Digi-NewB was conducted to propose methods to meet clinical monitoring needs [2]. Its goal was the continuous monitoring both to detect potential diseases and to evaluate the maturation of the newborn brain. The latter, related to neurodevelopmental outcome, is an important criteria for discharge home and is the focus of this paper.

Premature newborns, i.e., born before a Gestational Age (GA) of 37 weeks, are hospitalized in Neonatal Intensive Care Units (NICU) where their vital functions, such as cardiac, cerebral and respiratory activities, are monitored. In addition to electrophysiological signals, other types of information (e.g., motion, vocal) have been shown to be relevant to assess newborns' development and health [3]. But nowadays, they are rarely studied because they usually require visual observations in the presence of newborns.

Motion is one of the most studied behavioral components of preterm newborn. Most of the studies on newborns' motion relied on general movement assessment, which provides a qualitative motion analysis for cerebral palsy detection [4–6]. In the field of neonatal seizures (see [7] for review), motion quantification methods have been applied to represent the motion globally [8] or locally [9]. It was also shown useful in the context of sleep staging whose organization reflects the brain maturation [10, 11]. Indeed, sleep staging is mainly based on motion temporal organisation analysis (e.g., motion and non-motion periods characterization, in terms of periods number, duration). Motor activity information can be extracted from direct sensors (accelerometers [12], electromagnetic sensors [13] or pressure mattress [14]), or using video

cameras [8–10, 15, 16]. Among them, video has the advantage to be entirely non-invasive and contact-less. Thus, it limits the risk of infection and preserves the comfort of the newborns. However, videos need to be processed to retrieve the motion information. In early works, motion analysis was manually performed [10,17]. But visual assessment, besides being time consuming, is also expert-dependent. More recently, thanks to the improvement in video processing, automated methods have been proposed to characterize motion [5, 8, 9, 15, 16]. In these studies, authors first estimate motion activity as a temporal motion series [8, 9, 15] or as a velocity field [5, 16]. From there, amplitude [8, 15], duration [5, 9] or frequency [5] features are computed.

Nevertheless, on the one hand, these studies were carried out on short recordings, and, on the other hand, under non-realistic NICU conditions, making it impractical to integrate them into continuous monitoring tools. Challenges induced by such context are numerous as described in [18]. Some of them were recently tackled such as the automatic detection of adults in the camera field [15] or the detection of periods of sole presence of the newborn [19]. But these works are only components that still need to be combined to achieve a fully automatic characterisation of the motion of newborns adapted to a continuous monitoring.

Moreover, the evaluation of maturation based on either automatic or observational characterization of the motor activity during the whole hospitalization of preterm newborns has never been investigated in the literature while, as previously mentioned, its importance in the analysis of many pathologies has been underlined.

This led our team to propose a new approach to integrate motion analysis within the objective of non-invasive monitoring of the neurobehavioral development. We hypothesize that the motion temporal organization (duration of motion and non-motion intervals, numbers of these intervals) reflects the newborn maturity, such as the organization of sleep stages. Secondly, we suggest that these motion features can be used to estimate a functional age that should be in accordance to the true Post-Menstrual Age (PMA) of the newborn to consider that he/she is developing properly. For this purpose, contributions of this work consist of:

- The collection of videos of newborns with a wide range of GA and PMA, recorded in real context of NICU;
- The development of a fully automated process to characterize temporal organisation of the motion composed of motion estimation, motion segmentation (to retrieve motion and non-motion intervals) and motion features extraction;
- The fusion of motion features to build a unique indicator called "Functional Motion Age" (FMA).

This paper is organized as follows. Materials and methods are presented in Section 2. Section 3 is devoted to the validation of the motion features extraction process and the evaluation of the FMA estimation for monitoring temporal motion organization in preterm newborns. This paper ends with a Discussion (Section 4) and a Conclusion (Section 5).

2 Materials and methods

The global framework of our method to estimate and characterize the motion of newborns from video recordings of long duration is depicted in Figure 1. It is composed of five steps, which are described in this section: motion estimation, detection of sole presence of the newborn, motion segmentation to retrieve motion and non-motion intervals, motion features extraction and estimation of the FMA.

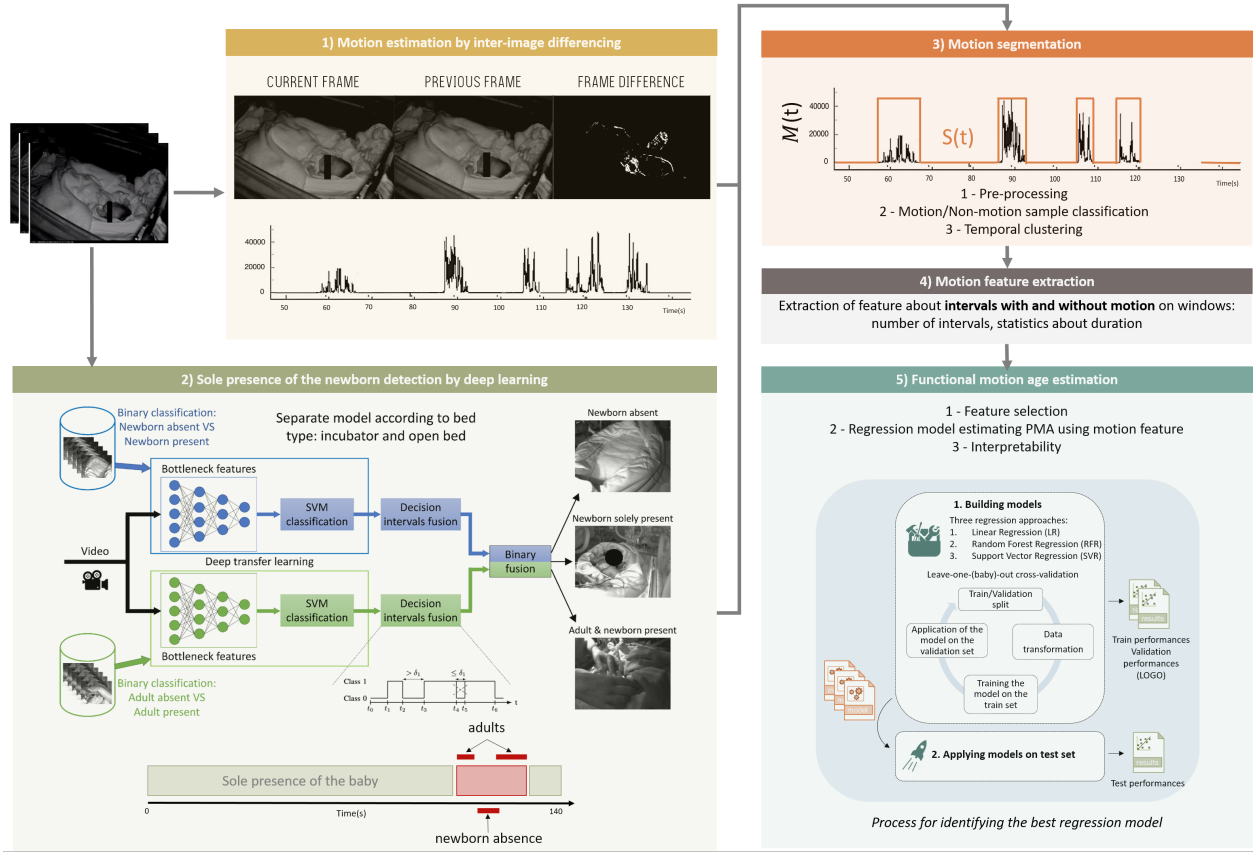


Figure 1: Overview of the proposed approach to characterize motion from videos in newborns: 1) motion estimation, 2) detection of sole presence of the newborn, 3) motion segmentation to retrieve motion and non-motion intervals, 4) motion features extraction and 5) estimation of the FMA.

2.1 Data acquisition

Video recordings were acquired in the scope of the Digi-NewB project [2]. This study received ethics approval from the Ouest IV Ethics Committee (reference number 34/16) and one of the parents of each newborn gave its signed agreement to take part to the study. Nineteen acquisition devices were deployed in six French hospitals. Acquisition devices embedded two near-infrared cameras with a resolution of 752x480 pixels (FMVU-03MTM-CS), which provide grayscale images. Video streams were recorded at $F_s = 25$ frames per second with MPEG-4 encoding. Cameras were placed between 30 and 80 centimetres of the infant head and were oriented in order to see newborn entirely at the feet and inside for incubators, at the head and outside for radiant warmers and cradles (see [18] for a more detailed description). In this work, data from a single camera per recording were studied.

2.2 Motion estimation

In literature, three classes of methods can be distinguished for motion estimation in newborns: inter-image difference, optical flow and block matching [3]. In this study, we focus on inter-image difference since it allows a global quantification of motion, which is sufficient for the characterization of the temporal motion organization. In addition to the fact that the distance between the newborn and the camera is changeable, newborns are usually very covered, in a very dark environment, making more sophisticated approaches oversized. Thus, raw motion series $M(t)$ are built from the number of changing pixels between consecutive frames [11, 15], as follows:

$$M(t) = Card(I(t,p) - I(t - 1/F_s, p) > T_h) \tag{1}$$

where F_s corresponds to the video frame rate, $I(t,p)$ is the intensity of the pixel p at current time t, $I(t-1/F_s, p)$ is the intensity of the pixel p at previous frame. T_h is a change in intensity threshold, set to 10, which limits the influence of noise related to the sensitivity of the camera on $M(t)$. This value was defined by studying histograms of change in intensity between images of videos of empty rooms, which is equivalent in decision theory to adapt the threshold under the hypothesis of no event.

2.3 Detection of sole presence of the newborn

The aim of this step is to discard unusable periods of video due to adults' interventions (i.e., adults in the camera field and absence of the baby from the bed) and to focus on periods of sole presence of the newborn in the bed. This point has been less investigated than motion estimation but some recent papers proposed to automatically detect periods of sole presence of the baby with deep learning techniques [19].

In this paper, we use a method recently proposed by our team, which relies on the classification of still images using deep transfer learning [20]. Figure 1 gives an overview of the method. Two binary classifications are performed: newborn absent vs present, and adult absent vs present. Bottleneck features are extracted from a neural network pre-trained with ImageNet [21] and are the input of a support vector machine classifier, which outputs a binary decision. Then, a temporal smoothing of the binary decisions is performed. Lastly, the two binary decisions are fused in order to output a three-class decision: adult & newborn present, newborn absent and newborn solely present. Finally, it provides a temporal indication of when the baby is present and alone in the video. This allows to replace values corresponding to not exploitable periods of the motion series $M(t)$ by missing values.

A separate model was created according to bed type: incubator and open bed. For incubator, the pre-trained neural network is MobileNetV2 [22] for "newborn" classification and InceptionV3 [23] for "adult" classification. In open bed, the pre-trained neural network is VGG16 [24] for both "newborn" classification and "adult" classification.

2.4 Motion segmentation

First, a ground truth had to be built from manually annotated data. Then, the proposed automatic motion segmentation relies on a three-step strategy. First, a pre-processing is applied to clean the motion series from noisy components. Secondly, a motion/non-motion classification is performed on each frame. Thirdly, a clustering is used to merge and discard short motion and non-motion periods. Strategy for optimizing these steps, according to ground truth, is given at the end.

2.4.1 Motion/Non-motion annotations

A manual motion segmentation needed to be done to obtain a ground truth. For this purpose, an expert visualized a subset of videos and labelled motion intervals using ViSiAnnoT (Video Signal Annotation Tool). It is an open-source Python package, developed by our team, that provides a graphical user interface for the visualization and annotation of video and signal data [25]. The instructions about installation may be found at the following link: `https://pypi.org/project/visiannot`. From there, a motion/non-motion reference segmentation was built by associating a value of 0 for a sample belonging to a non-motion interval, 1 otherwise.

2.4.2 Pre-processing

Two pre-processings are applied to get clean motion series $M_c(t)$: median filtering and correction of flashing artifacts.

Median filter: A median filter of size s was used to remove artifacts resulting from sudden light variations (e.g., switching on/off) that induce localised high-amplitude peaks in the motion signal $M(t)$.

Correction of flashing artifacts: The second pre-processing is related to the photo-detector used to measure blood oxygen saturation (also called pulse oximetry). Although the flashing is invisible to the naked eye, it is captured by the infrared camera. Its impact varies from one recording to another, but also throughout a single video. In fact, when the detector is placed on one of the newborn's foot or finger, it may be hidden by the blanket and have no impact on motion series. Reversely, when the baby moves and the photo-detector becomes discovered, a noise impacts $M(t)$ more or less significantly depending on the reflection of the flashing on the blanket. An example is presented in Figure 2.

To reduce the impact of artifacts, we first apply a n-order Butterworth filter. Two filtering strategies were investigated: lowpass with the cut-off frequency f_l and band-stop with f_{b1} and f_{b2} frequencies. Then, we use the Baseline Estimation and Denoising with Sparsity (BEADS) algorithm [26] to retrieve a flat baseline. This method is based on the hypothesis that the vector of observations $y(t)$ can be modelled as:

$$y(t) = x(t) + f(t) + w(t) \tag{2}$$

where $x(t)$ is a sparse signal, $f(t)$ is a low-pass baseline and $w(t)$ is a stationary white Gaussian noise. The sparse signal $x(t)$ is based on two parameters: the asymmetric ratio of peaks r and a regularization parameter amp. Baseline $f(t)$ is adjusted by two parameters: the filter order d and the filter cut-off frequency f_c. As the shape of our signals is similar to that studied by the authors of the BEADS algorithm, we use the parameter values they proposed, as follows: $d = 1$, $f_c = 0.01$, $r = 6$, $amp = 1$. The source codes provided by the authors are available at:`https://github.com/hsiaocy/Beads`.

2.4.3 Motion/Non-motion sample classification

The aim of this step is to define a set of features for motion and non-motion binary classification in order to get an automatic labelling for each sample. In the literature, several groups tackled pathology classification (e.g., cerebral palsy, neonatal seizure) using features (e.g., trajectories, speed or acceleration) computed

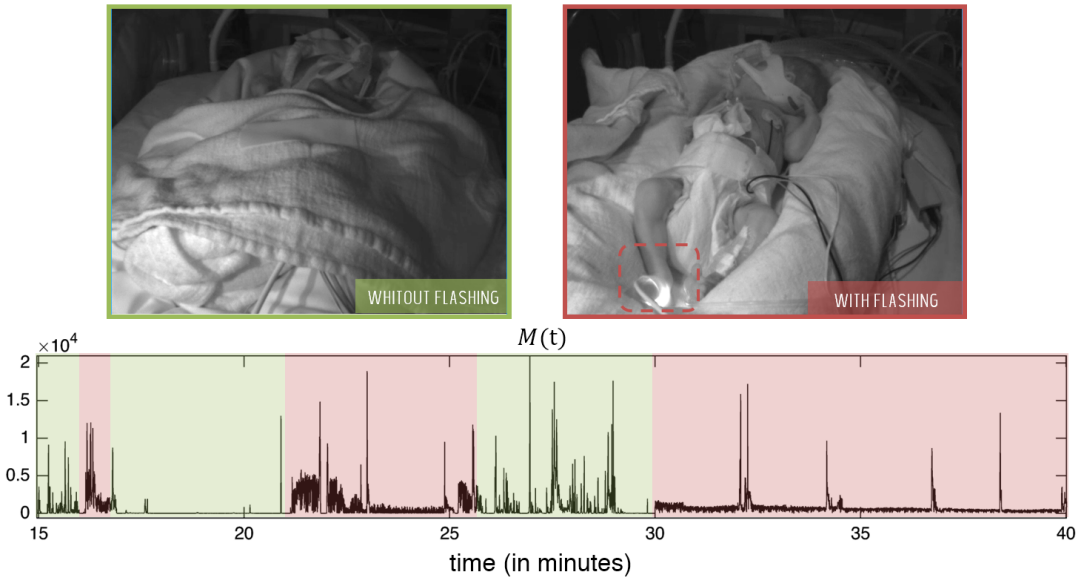


Figure 2: Example of a motion series $M(t)$ without (in green) and with pulse oximetry flashing (in red) with associated images.

from motion [9, 27, 28]. However, these features are not appropriate for a motion/non-motion classification of sample and difficult to compute when newborns are covered with a blanket.

We therefore propose to work with a feature set that has been proven to be effective in a similar problem, namely electroencephalographic (EEG) burst detection [29]. Features are of three types (amplitude, statistical, energy) and computed on windows of w seconds centred on each analysed sample. They are reported in Table 1. These features are used as an input to classification algorithms in order to label each sample as motion ("1") or non-motion ("0"). We investigated the following algorithms: Logistic Regression (LogR), K-Nearest Neighbors (KNN) and Random Forest (RF).

2.4.4 Temporal clustering

A final temporal clustering is performed in order to get a motion segmentation signal $S(t)$. If a motion interval appears less than five seconds after another, the two intervals are merged. Then, all motion intervals that last less than two seconds are eliminated (see Figure 3). These values were defined on the basis of literature values [8], as well as from the experience acquired during scoring.

2.4.5 Tuning of pre-processing parameters and hyper-parameters of classification models

For each of these steps, parameters and hyper-parameters are optimised with the aim of maximizing the F1-score. The choice of this metric is supported by the fact that the number of non-motion samples is far greater than the number of motion samples. Grid search and Leave-One-Group-Out (LOGO) cross-validation, where a group is defined as the set of recordings of one given baby (i.e., because several recordings of a same newborn can be part of a dataset), was performed in order to identify best hyper-parameters. The summary of the tuning of these parameters is presented in Section 3.3.1.

Table 1: Input features for motion/non-motion classification. Each feature is computed on a window of w seconds with $n = w.F_s$ samples ($F_s = 25$).

Feature	Description
M_m, M_{m-1}, M_{m+1}	Differences in amplitude between the maximum and minimum motion values in the current, previous and next window respectively.
DM	Maximum of absolute motion values of the first order derivative $DM = \max_{k=1,\ldots,n} \{\mid M_c(k) - M_c(k-1) \mid\}$
Sd	Standard deviation
Kt	Kurtosis
NL	Non Linear Energy Operator [30] $NL = \frac{1}{n} \sum_{k=1}^{n} M_c(k)M_c(k-3) - M_c(k-1)M_c(k-2)$
RMS	Root Mean Square
AD	Averaged differentiation $AD = \frac{1}{n} \sum_{k=1}^{n} \mid M_c(k) - M_c(k-1) \mid$

Figure 3: Illustration of the temporal clustering step: motion intervals occurring within 5 seconds of each other are merged and motion intervals of less than 2 seconds are deleted.

2.5 Motion features extraction

Once the motion segmentation is performed, we characterize the intervals by different features. These features were proposed in order to describe motion and non-motion intervals in terms of duration and number. The whole features set is reported in Table 2. It has the particularity of being independent of the amplitude, which is influenced by external factors like the baby size or the recording conditions (e.g., shadows, zoom, blanket).

Table 2: Definition and naming of the motion set of features.

Feature description	From motion intervals	From non-motion intervals
Total duration (ratios)	TotalWithMotion	TotalWithoutMotion
Mean duration (seconds)	MeanWithMotion	MeanWithoutMotion
Median duration (seconds)	MedianWithMotion	MedianWithoutMotion
Maximum duration (seconds)	MaxWithMotion	MaxWithoutMotion
Standard deviation of duration (seconds)	SDWithMotion	SDWithoutMotion
Relative standard deviation of duration (seconds)	RSDWithMotion	RSDWithoutMotion
Number of intervals (scaled to an hour)	NBWithMotion	NBWithoutMotion

These features are computed on sliding windows of 5 minutes with 50% of overlap. Additionally, in order to work with relevant features, only windows with more than 75% of sole presence of the newborn are kept. This last constraint, in addition to the remaining missing values, can lead to a non-complementary between motion and non-motion features, which justifies that we study all of them.

In case of a long duration recording, features of the first 5 hours of exploitable data (i.e., sole presence of the newborn) are kept. At the end, each recording was summarized by the medians of the features meaning that 14 features were obtained on these 5 exploitable hours. This approach, although strict, is a precaution taken to ensure that motion features from different recordings will be comparable to each other while keeping enough records to build.

2.6 Functional motion age estimation

In order to build an estimator of the FMA, regression approaches have been investigated to fuse motion features. For that purpose, models were trained by targeting PMA (in weeks). In this section, we present our approach for selecting features, training and evaluating regression models.

2.6.1 Feature selection

An initial feature selection is made. In fact, too many features and particularly highly correlated ones can bring noise in models and/or may lead to overfitting. When high Pearson's correlation (>0.95) are obtained between two features, only one is kept (i.e., the one with the highest Pearson's correlation with the target).

2.6.2 Evaluation and training of regression models

Three regression approaches were investigated: Linear Regression (LR), Random Forest Regression (RFR) and Support Vector machine Regression (SVR).

LOGO cross-validation was performed in order to compare model and to identify best hyper-parameters of models. The process of training and evaluating the model is performed several times as follows (see Figure1):

- Train and validation split: data of one newborn are kept for the validation, all others are used in the training set.
- Data transformation: each numerical variable of the training set is converted into its Z-score (i.e., standardization).
- Training a RFR, SVR or LR model: this is performed on the training set, using a current set of hyper-parameters.
- Application of the resulting model on the validation set.

At the end, a final training is made on the entire dataset for an application of models on a test set. The finally used hyper-parameters were the ones leading the lowest Mean Absolute Error (MAE) during LOGO. This metric allows to easily interpret the errors as they are directly expressed in weeks. The summary of the tested hyper-parameters using grid search is given in Section 3.4.1.

2.6.3 Interpretability

The importance of each feature in regression models was also studied. We retrieved coefficient related to each feature for LR. To compare, feature importance scores were retrieved for RFR and SVR models. Although intrinsically present in RFR models, this information is not directly available for SVR because a transformation is first applied to the feature set. Therefore, to obtain the importance of the motion features, we applied a model-agnostic method: permutation feature importance [31].

From there, to enhance interpretability, we intend to investigate the most compact model using a minimal number of features. For LR, features with the highest regression coefficient will be kept. For RFR and SVR, the most important features will be defined as the ones with a permutation importance greater than 0.10.

Then, we proposed to look dynamics of the features of used in the best regression strategy. It was built by processing videos of a control population to get one motion feature set by recording. In a second step, data of the entire population were clustered in two-week PMA intervals. From there, mean value and 95% confidence interval were calculated among all the newborns for each of these intervals and for each feature. This resulted in expected dynamics of each motion feature.

3 Results

3.1 Video datasets

In order to evaluate and study the relevance of our approach, two datasets were extracted from the Digi-NewB database: a Segmentation Dataset (SD) and an Healthy Preterm Dataset (HPD). The first one was built to evaluate the motion segmentation method and the accuracy of extracted motion features, and manually annotated. The second one was constructed to evaluate FMA estimation and discuss the clinical relevance of the motion feature set.

3.1.1 Segmentation dataset

SD is composed of video recordings of 33 newborns with a high range of GA (from 25.3 to 41.6 weeks). A set of 57 recordings of 30 minutes was selected. Video data are selected for each of the six hospitals of the Digi-NewB project. In order to challenge our segmentation method, balance was forced between recordings with (29 recordings) and without (28 recordings) flashing throughout the recording during the selection. In real life, flashing only appear at some periods in the recording and its presence is not known by advance. For each category, a particular attention has been paid to the distribution of recordings in incubator or open

bed. The expert annotated each video recording of the dataset for a total duration of 1,710 minutes and a wide range of included PMA. The dataset description is summarized in Table 3 and the distribution of the GA and PMA ages are shown in the Figure 4. We can observe that GAs are more concentrated in the range [26-32 weeks]. This is related to the choice to select babies who were passing through different environments, thus those with the lowest GAs. In addition, a gap between 32 and 34 weeks PMA can be observed. It is due to the balance between recordings with flashing and without flashing and to the balance between recordings in incubator and open beds that we intended to capture.

Table 3: Clinical content and recording settings of SD (segmentation dataset).

CLINICAL/RECORDING	DATA
Number of newborns	33
Gender	14 females and 19 males
Hospital id (number of babies)	1 (16), 2 (1), 3 (8), 4 (3), 5 (4), 6 (1)
Number of recording sessions	57 (33 in incubators)
Duration per recording	30 minutes
Flashing in recordings	29 with (58.6% in incubators) and 28 without (53.6% in incubators)

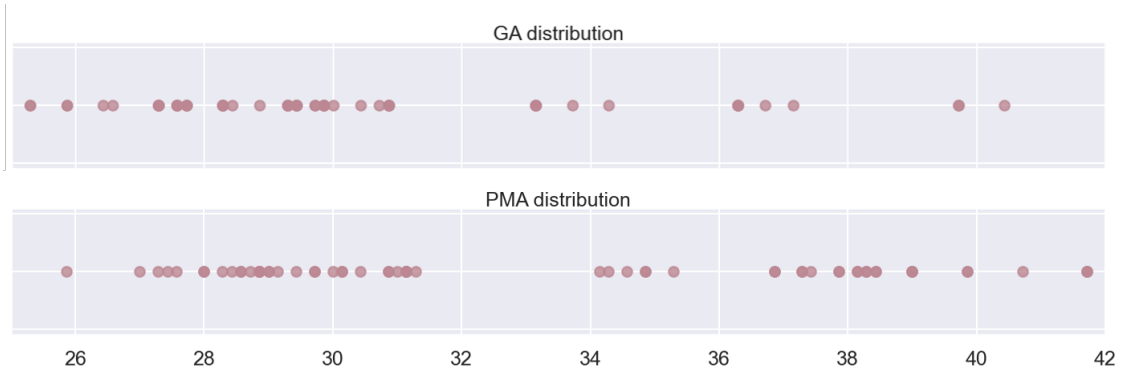


Figure 4: Post-Menstrual Age (PMA) and Gestational Age (GA) distribution (in weeks) in the Segmentation Dataset.

3.1.2 Healthy preterm dataset

HPD includes video data from 46 preterm newborns with a high range of GA (from 25.6 weeks to 41.6 weeks), corresponding to a total of 205 recordings available in the Digi-NewB database with a PMA range between 26.1 and 41.7 weeks. Recordings last 24 hours each. This population has been retrieved after a careful examination of medical records by two neonatologists. They identified preterm newborn without any of the following exclusion criteria: chest compression for resuscitation at birth; severe neurological lesions (intraventricular haemorrhage of grade 3 or 4, white matter lesions, hypoxic-ischemic encephalopathy); early or late onset sepsis; enterocolitis; severe malformations; and preterm infants with a birth weight lower than the 10th percentile for their GA. Secondly, they verified that infants presented trajectories during the entire period of observation that could be considered normal for their GA. Finally, this population was randomly divided into two parts: 2/3 for training and 1/3 for testing (see Table 4 and Figure 5).

Table 4: Clinical content and recording settings of HPD (Healthy Preterm Dataset).

CLINICAL/RECORDING	DATA (train-validation)	DATA (test)
Number of preterm newborns	31	15
Gender	9 females and 23 males	8 females and 7 males
Hospital id (number of babies)	1 (17), 3 (8), 4 (1), 6 (5)	1 (8), 3 (3), 4 (1), 6 (3)
Number of recording sessions	113 (41 incubators)	92 (61 incubators)
Duration per recording	24 hours	24 hours

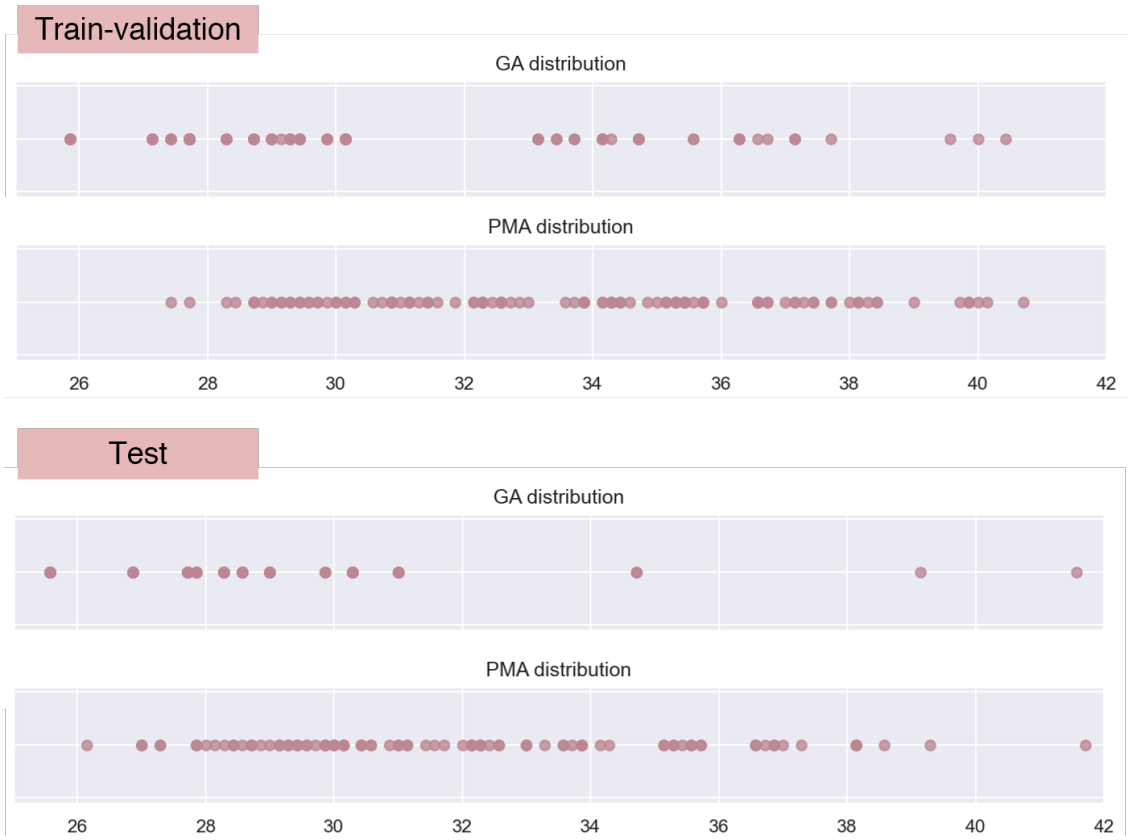


Figure 5: Post-Menstrual Age (PMA) and Gestational Age (GA) distribution (in weeks) in the Healthy Preterm Dataset.

The selection of newborns by neonatologists was focused on the care pathways and not on the search for a balance in terms of age. We can therefore note that GA are not well balanced, especially on the test set. However, PMA ages are much better represented for both train-validation and test sets. To verify that the random selection does not create a divergence of distribution between train and test sets, we performed Kolmogorov-Smirnov tests [32] between GA of the train set and GA of test set and between PMA of the train set and PMA of the test set. No evidence of belonging (i.e., p-values < 0.05) to different distributions was revealed (p-value=0.19 and p-value=0.34, respectively).

3.2 Software and platforms

Experiments presented in this study were performed using Python 3.7.3. Machine learning developments were made with the support of scikit-learn 0.23.2 and statistical analyses were conducted using scipy.stats 1.6.0.

3.3 Evaluation of the motion characterization

In this section, the motion segmentation approach is studied. The tuning of the pre-processing parameters and hyper-parameters is firstly described. Then, methods are evaluated from two angles: sample by sample performance of each classifier and accuracy of the resulting extracted motion features. Experiments were conducted on the Segmentation Dataset.

3.3.1 Tuning of pre-processing parameters and hyper-parameters of motion/non-motion classification models

Parameters were tuned for each of the three classification approaches (KNN, LogR and RF). In each case, LOGO was performed. Parameters at different levels of the segmentation method were optimised, such as the window size of the median filter, the frequencies and filter types of the baseline modelling, the window over which features are calculated for classification, and the hyper-parameters of the three learning approaches. Table 5 summarizes the tests conducted for each parameter.

Table 5: Summary of the tuning of pre-processing parameters and classifiers hyper-parameters for motion/non-motion classification. KNN stands for K-Nearest Neighbors, LogR stands for Logistic Regression and RF stands for Random Forest. For each classifier, we give the values selected by parameters tuning.

Method	Parameter	Tested values	Selected values		
			KNN	LogR	RF
Median filtering	s	1, 5, 7, 11	5	11	7
Butterworth filtering	Lowpass f_l	1.5, 3, 6, 8.5, 10	-	8.5	-
	Bandstop $[f_{b1}, f_{b2}]$	[1.5, 3.5], [1.5, 6.5], [1.5, 9], [1.5, 11]	[1.5, 6.5]	-	[1.5, 3.5]
Features extraction	w	1, 2, 5	1	2	2
KNN classification	$n_{neighbors}$	1, 5, 10	1	-	-
	weights	uniform, distance	distance	-	-
	p	1, 2 (when weights = distance)	1	-	-
LogR classification	c	10e-3, 10e-2, 10e-1, 1, 10	-	10e-2	-
RF classification	$n_{estimators}$	1, 5, 10, 50	-	-	50
	criterion	entropy, gini	-	-	gini

Final parameters (i.e., the ones that led to the highest F1-scores) are identified for each approach. No

global set of parameters for pre-processing steps emerges since they revealed to be dependent on the studied classification method.

3.3.2 Sample by sample performance of classifiers

The best segmentation performances for each classification approach are compiled in Figure 6. Distributions of F1-scores over the cross-validation are reported without and with applying pre-processing steps.

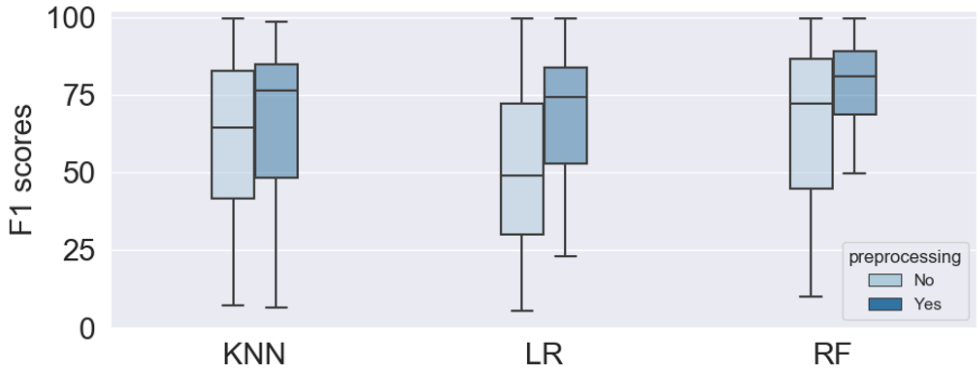


Figure 6: Boxplots of F1-scores without and with pre-processing on validation sets over the cross-validations for the three classification approaches: K-Nearest Neighbors (KNN), Logistic Regression (LogR) and Random Forest (RF).

Median F1-scores of classifiers are higher when pre-processing steps are applied for the three types of classifiers. They increase from 65 to 77% for KNN, from 49 to 75% for LogR and from 72 to 81% for RF. Overall, best results are obtained with pre-processing associated with the RF approach. F1-scores range between 50% and 99% with the 25th percentile at 70% and the 75th percentile at 90%. The performances of KNN and LogR are weaker and much less stable over cross-validations.

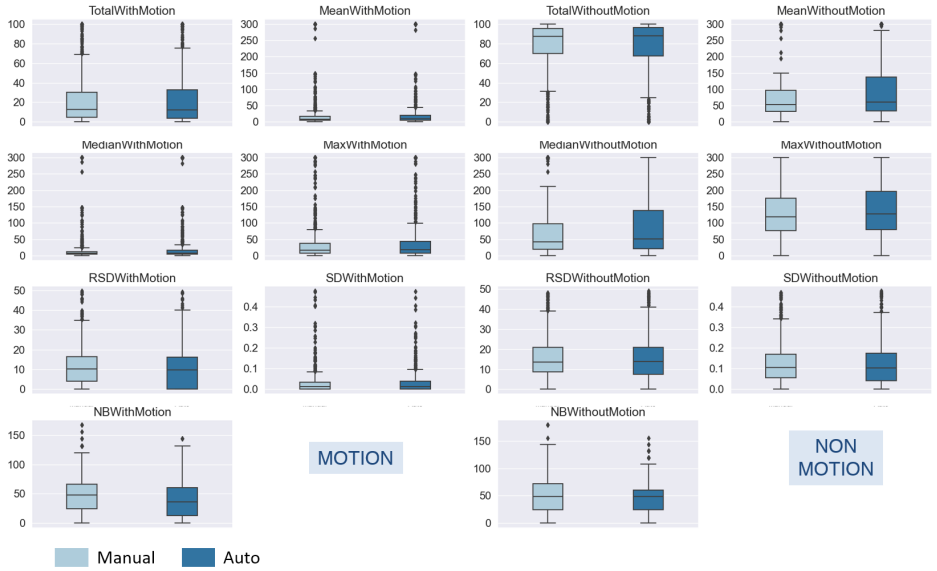
In addition, an average precision of 77% and a recall of 78% were obtained with the RF model.

The similarity of these two measures indicates that despite the moderate performance, the classifier is as responsive to the classification of motion samples as it is to the classification of non-motion sample.

3.3.3 Evaluation of the accuracy of the extracted motion features

To evaluate the accuracy of resulting motion features, features estimated using the automatic segmentation given by the RF model were compared with the ones computed from the reference segmentation. For all the manually annotated recordings, medians along recording of the fourteen features were computed from the values obtained on sliding windows of 5 minutes with 50% of overlap as described in Section 2.5. Distributions of features values given by automatic and manual segmentation are provided in Figure 7.

As a first observation, the distributions of feature values obtained from the automatic segmentation are similar to those obtained from the manual segmentation. To verify this statement statistically, equivalence tests (i.e., two-one-sided t-tests) were carried out. It allowed to verify the equivalence between values while accepting a margin of error [33]. Here, errors between -5 and 5 seconds for features expressed in seconds, -5% and 5% for features expressed in percentage and between -5 and 5 for features expressed in number, were admitted. Beforehand, the normality of the distributions was checked with Shapiro's test for each feature. If normality condition was not respected (p-values < 0.05), a Box-Cox transformation was applied



Motion Feature	Pearson's r
MeanWithMotion	0.9
TotalWithMotion	0.9
TotalWithoutMotion	0.9
MaxWithMotion	0.9
MedianWithMotion	0.9
NBWithoutMotion	0.8
NBWithMotion	0.8
MaxWithoutMotion	0.8
MeanWithoutMotion	0.8
SDWithMotion	0.8
MedianWithoutMotion	0.8
SDWithoutMotion	0.6
RSDWithoutMotion	0.6
RSDWithMotion	0.6

Figure 7: Boxplots of the distribution of the fourteen features obtained from automatic and manual segmentation. Reported values are in seconds except for NBWithMotion and NBWithoutMotion which are in number of units, RSDWithMotion, RSDWithoutMotion, TotalWithMotion and TotalWithoutMotion that are percentages. Pearson's correlation coefficients are reported on the right.

to normalize feature before performing the equivalence test [34]. At the end, the hypothesis of equivalence was not rejected (p-values < 0.05) for any feature.

Secondly, Pearson's correlation coefficients between automatically and manually calculated values of each feature were computed. They are reported in Figure 7. Mostly strong correlations (≥ 0.7) are reported except for SDWithMotion, RSDWithMotion and RSDWithoutMotion where moderate correlation coefficients of 0.6 can be noticed.

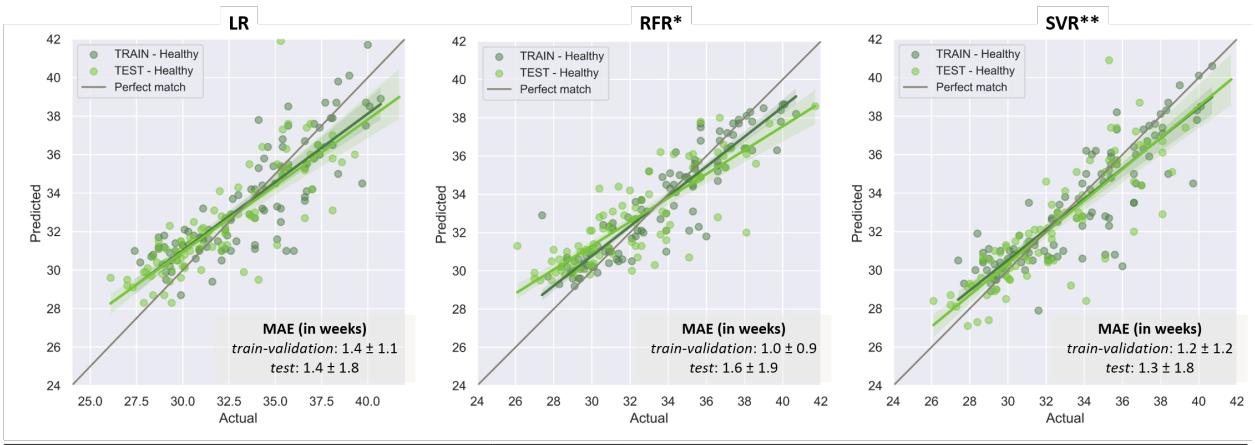
3.4 Evaluation of the functional motion age estimation

In this section, experiments were conducted on HPD. We apply and study the motion features fusion process within the objective to build the estimator of the FMA applicable to the entire extremely preterm population.

3.4.1 Performance of the modelling approaches

Modelling approaches were compared to estimate functional motion age on HPD. As presented in Section 2.5, each recording was summarized by 14 motion features. Then, they were used as input features along with GA for regression modelling. It was reduced to 10 features with Pearson's high cross-correlation filtering. Best hyper-parameters were identified using LOGO cross-validation on the train-validation set and finally used on this whole dataset for final training. The generalisation capacity of final models was then evaluated on the test set. Results are reported in Figure 8.

All regression approaches succeeded with an averaged MAE between 1.0 and 1.4 weeks to predict the true PMA whether on the train-validation set and between 1.3 and 1.6 weeks on the test set. The standard deviations are larger on test sets than on train sets, which denotes a tendency to overfit the training data



Feature	LR Coefficient	RFR Importance	SVR Importance	Feature	LR Coefficient	RFR importance	SVR Importance
GA	1,89 (2)	0,42 (1)	0,39 (2)	SDWithoutMotion	-1,22 (4)	0,03 (3)	0,03 (7)
TotalWithMotion	-0,27 (7)	0,01 (5)	0,14 (4)	RSDWithMotion	-0,13 (8)	0,01 (5)	0,04 (6)
MeanWithMotion	2,18 (1)	0,39 (2)	1,19 (1)	RSDWithoutMotion	1,32 (3)	0,02 (4)	0,07 (5)
MeanWithoutMotion	0,67 (5)	0,01 (5)	0,07 (5)	NBWithMotion	-0,05 (9)	0,02 (4)	0,01 (8)
MaxWithoutMotion	0,01 (10)	0,01 (5)	0,03 (7)	NBWithoutMotion	-0,35 (6)	0,02 (4)	0,20 (3)

*final hyper-parameters RFR identified during Leave-One-Group-Out cross-validation: number of decision trees = [**50**, 100, 200], min sample split = [1, 5, 11, **17**]
final hyper-parameters SVR identified during Leave-One-Group-Out cross-validation: kernel = [linear, **radial basis function, polynomial], C=[1000, **3000**, 10000], gamma = [**1e-3**, 1e-4]

Figure 8: Regression performances presented by correlation plots, Mean Absolute Error (MAE) and standard deviation for the three approaches: Linear Regression (LR), Random Forest Regression (RFR) and Support Vector machine Regression (SVR). Final regression coefficients in LR model and importance of features of RFR and SVR models are also presented along with their rank in the model. The tested hyper-parameters values for SVR* and RFR** approaches with final selected values (in bold) are reported at the bottom.

for all three approaches. This can be particularly noticed for RFR with a larger gap between the average MAE on the train-validation and on the test sets. To produce the FMA, SVR seems to be the more stable approach with a MAE of 1.2±1.2 weeks on train-validation and 1.3±1.8 weeks on test. This is confirmed by the SVR correlations plot, with a line regression on the train-validation that coincides almost perfectly with that obtained on the test. The LR approach also reveals suitable results. For all approaches, there is a tendency to underestimate age when it is higher (above 34 weeks).

3.4.2 Interpretability

Regarding the importance of each feature (compiled in Figure 8), we can notice that mean duration of motion intervals (MeanWithMotion) has a rather high importance for all models. It has the highest coefficient (2.18) for LR, an importance of 0.39 for RFR and an importance of 1.19 with SVR. It should be noted that it was one with the highest Pearson's correlation coefficient obtained by our automatic segmentation (see section 3.3.3). GA revealed important for prediction for all approaches. For SVR, NBWithMotion and TotalWithMotion are also of significant importance with values of 0.20 and 0.14 respectively. Fortunately, features that showed the lowest Pearson's correlation coefficients during the evaluation of the motion characterization (SDWithoutMotion, RSDWithMotion and RSDWithoutMotion) has only a weak importance on predictions in SVR and RFR model. As this is not the case with LR, it decreases the reliability of this modelling. At the end, SVR appears to be the best strategy.

To complete the study, we performed the experiments again using only the most influencing features of SVR (GA, MeanWithMotion, TotalWithMotion and NumberWithMotion) in order to investigate a more compact model. This results in a MAE of 1.2±1.2 weeks in train-validation and 1.3±1.2 weeks in test. In other words, the use of less features reduces the variability of our predictions on the test set and reinforces the generalization of our approach while increasing the interpretability of the predictions process.

Dynamics of its three most important features, beside GA, are given in Figure 9. As a first observation, a

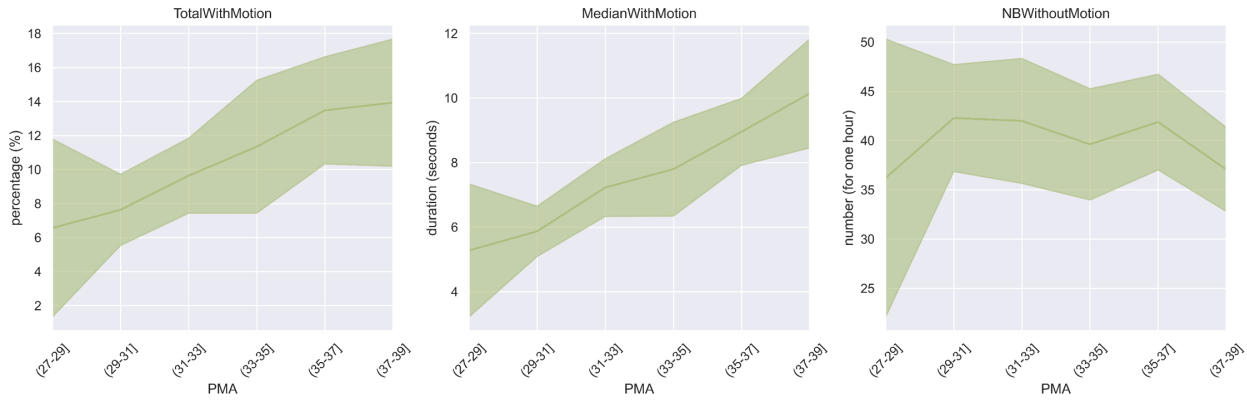


Figure 9: Evolution with PMA in premature newborns of the three most important features of the SVR model.

tendency to increase can be observed for motion interval features (MeanWithMotion and TotalWithMotion). Dynamic of the number of intervals without motion (NBWithoutMotion) is less well defined but a slight trend in the decrease of variability with PMA can be remarked. Regarding values of features, MeanWithMotion lasts in between 4 and 16 seconds and raises from 6 to 13 seconds. In order to interpret these values properly, it should be remembered that they refer to a 5-minute window analysis. One can also notice that newborns generally spend more time not moving than moving (e.g., see percentages of TotalWithMotion that is scaled

between 0% and 18%). This percentage doubles between the lowest and highest ages. For its part, the number of intervals without motion is comprised between 22 and 50 in the 27-29 PMA range and between 33 and 42 in the 37-39 PMA range.

4 Discussion

In this work, a fully-automated process for characterizing motion temporal organisation of preterm newborn from long-term video has been designed and evaluated on real-life conditions. For the first time, 4,948 hours of video data of premature newborns have been automatically processed. We have covered a wide range of PMA (from 25.3 to 41.8 weeks) and in several hospitals, which implies a large number of room configurations (incubator, radiant warmer and cradle).

Some parts of the proposed processing chain have already been validated in previous works such as the motion estimation [15] or the automatic detection of sole presence of the newborn [20]. For the rest, although methods for characterising motion have been proposed [15, 35], they had not been evaluated under real conditions. Here, a new method for motion segmentation, a new set of features and a novel way of combining them to estimate a FMA were proposed.

Firstly, a segmentation method to retrieve motion and non-motion intervals was developed. This approach takes into account artifacts such as lighting variation and pulse oximetry by integrating pre-processing of the raw motion signal. Then, from a previously validated feature set for EEG burst detection [29], machine learning techniques were investigated. This yielded one main result: Random Forest model can be trained to accurately characterize motion and non-motion intervals. Indeed, no major statistical difference was observed between the motion features extracted from the automatic segmentation and those issued from the manual annotation.

Secondly, a new motion feature set focusing on motion and non-motion intervals' statistics on duration and intervals counting was defined. Its main strength is its robustness with respect to acquisition conditions (e.g., distance to the camera and orientation) contrary to amplitude features [15].

Then, we proposed to aggregate these motion features into a unique indicator that we have called FMA. For that purpose, we suggested to estimate PMA from these features since they are also supposed to evolve with age. Three regression approaches were studied: Linear Regression, Random Forest Regression and Support Vector machine Regression. Results have revealed that it was achievable with an averaged mean absolute error of 1.3 weeks using a SVR model. This value is not surprising when looking at confidence intervals observed in dynamics of separate features which are quite wide. To our knowledge, this is the first attempt to propose such an indicator for clinical decision support. It could meet the need of clinicians to have at hand indicators allowing the continuous monitoring of the motor activity of the newborn.

Moreover, this indicator has the merit of being interpretable since a minimal set of features was identified (i.e., GA, MeanWithMotion, TotalWithMotion and NBWithoutMotion) and features dynamics has been studied. Indeed, this goes alongside the strong desire of clinicians to see the proposed new systems be interpretable or explainable for an effective transfer to care routine [36]. When applied to premature data, we saw that the most important features of the SVR model evolved with PMA. This is the first time that these observations have been made on such a large population and in an automatic way. In particular, we have seen that the percentage of time spent in motion increases with age. The results presented are a first attempt to characterize the maturation through the motion temporal organization in a population of premature newborns having had an evolution without severe complications identified during hospitalization. The use of machine learning algorithms is justified by the complexity of changes and the multiple determinants in maturation of temporal organization of the preterm infants spontaneous motion. Indeed, this organization depends on changes in sleep and wakefulness states and on the maturation changes of spontaneous movements during those states. It is known that time and average duration of time intervals spent in quiet sleep, which

corresponds to sleep without motion, increases with maturation [25]. Recently, it was also shown a decrease in the occurrence of motion intervals shorter than 30 seconds and an increase in motion intervals longer than 30 seconds [37]. Finally, it is known that time spent in wakefulness increases with maturation and that time spent in active sleep decreases during neonatal maturation of preterm infants [38] but without description of motion temporal organization during these periods. Our results are in agreement with these observations. However, due to their novelty, they require further study.

Some limitations and perspectives can be expressed.

In this study, the estimation of the motion series was done by an inter-frame difference, which is a basic approach, but sufficient to extract motion and non-motion periods. Other methods such as block matching, optical flow have been also studied in the literature. The difficult context of continuous monitoring did not allow us to consider the use of block matching, in particular because the babies are very covered which makes impossible to identify the regions of interest to follow. The use of optical flow methods may be further investigated, even if preliminary experiments have shown no benefit while being more computationally expensive.

As part of the segmentation approach, the frame per frame evaluation gave a quite wide F1-score range over cross-validation. This means that the generalisation of our motion/non-motion segmentation model will have to be studied in more details. To improve performance, it may be relevant to put more efforts on the tuning of parameters and hyper-parameters, especially the parameters used for the BEADS pre-processing. To date, we only applied parameters values reported by the authors which were optimized on chromatography [26]. Another point of improvement would also be to study in more details the impact of features for non-motion classification on our segmentation performance. One more point of attention concerns the annotation of the data for training. Some recordings with very little time spent in motion and with very short intervals of very low amplitudes were selected. However, the expert found it difficult to annotate them, which raises questions about what should or should not be considered as motion. Indeed, now that we showed that temporal motion organization is a potential maturation indicator, efforts should be directed to increase knowledge on this subject. For example, by defining what is to be considered as a relevant motion interval to evaluate maturation.

From the point of view of the extracted features, we can regret their simplicity which does not take into account the complexity of the motion carried out by the newborn. It would now be interesting to look inside the intervals of motion to extract information of a frequency [16] or trajectory [39] type. It could also be relevant to automatically divide the image into several zones in order to know the origin of the movements (arms or legs). Recent works based on deep learning computer vision techniques were proposed in that sense such as the automatic detection of newborns body outside the blanket [40] or the detection of limbs in fully uncovered newborns [6, 41]. However, these methods have been developed on restricted conditions (uncovered and/or daylight illumination) since their objective is not to ensure continuous monitoring. To obtain efficient models it will first require a significant amount of time to annotate real-world dataset for this specific task.

Increasing the features set may also help to provide a more accurate FMA. However, it is a major concern to keep in mind the interpretability of the proposed methods, which goes hand in hand with the sparing use of features, in order to be accepted by the clinicians [36]. It is for this need for interpretability that we chose not to adopt an end-to-end deep learning strategy, but rather to use deep modelling sparingly to address challenges that do not require clinical interpretation (e.g., detecting periods when the baby is alone).

With regard to the FMA estimation, a final MAE of 1.3 with a standard deviation of 1.2 weeks is reported on the test set. It may seem inaccurate to have an overestimation of PMA of 2.5 weeks. Some elements allow us to say that the prediction is still admissible. On the one hand, the observation of the dynamics of the parameters shows overlapping values when observed in intervals of 2 weeks. In another hand, although it has not been studied through motion characterization in the literature, maturation is known to be a rather slow process that does not take place at exactly the same age for all children, thus leading to an overlap in

the observations. To illustrate this, we can draw a parallel with the fact that Einspieler et al. observed the appearance of fidgety movements in a period of 3 to 6 weeks at corrected age, giving a 3-weeks overlap [42].

Other regression strategies can be also studied. Indeed, if the SVR model seems sufficiently stable and adapted, the evolution of the different parameters indicates mostly linear dynamics. This is not well taken into account by this approach. We could investigate an ensemble model (e.g., a linear combination with a non-linear model) for a better modelling.

Another avenue of investigation concerns the granularity of the analysis performed. Indeed, in this study, we estimated FMA with a set of parameters that represents 5 hours of data within a recording. We could imagine to decrease the granularity and keeping the same approach or by using recurrent network approaches because they would help to contextualize the successive predictions, as in survival neural networks [43]. However, in order to do so, the integration of non-exploitable periods will need to be carefully considered. Also, since maturation is a slow phenomenon, the granularity needed to correctly capture it would have to be studied. A good basis could be to study periods of at least 3 consecutive hours since this is the time of a complete sleep cycle in preterms [44].

The impact of the use of different acquisition systems (camera model, type of image, positioning, etc) on the performance of the method can also be discussed. Indeed, to answer this question, additional experiments should be carried out. This should have no impact on the motion estimation and the motion segmentation results provided that the camera does not have a higher noise sensitivity than our camera model (the threshold T_h should otherwise be increased) and that the camera is placed between 30 and 80 cm from the newborn. Concerning the detection of sole presence of the newborn, we can expect a good generalization if the images are grayscale images and transformed to the same dimensions. Indeed, the model, being designed by transfer learning, integrates the knowledge of various images. If poor performances are observed anyway, a fine-tuning on a smaller annotated dataset would be necessary. The rest of the chain, i.e., the estimation of the motion features as well as the estimation of the functional age, should not be impacted if the previous steps are efficient. Furthermore, our method should work with a higher or lower video frame rate, if enough samples are still available on the 2-second windows used to conduct the motion segmentation using the random forest model.

To date, limits in terms of clinical application can be raised. In this study, we have not used this indicator to discriminate between pathological and non-pathological cases. Indeed, this was a preliminary step, allowing us to verify that information related to maturation could be captured through motion analysis in a non-pathological population. In future works, we aim to apply it to two clinical situations: early detection of sepsis since it has been shown to be associated with an atonic behavior [45] and early detection of neurological disorders which manifests itself in motion disorders [46].

Clearly, this indicator alone is not sufficient for clinicians to accurately assess the development of the newborn. We believe that it is part of a whole and we are working on similar indicators automatically computed under real conditions and obtained from cardio-respiratory [47], vocal [48, 49] and sleep activities [11, 25].

5 Conclusion

In this paper, a fully automated process for characterizing newborns' motion temporal organization from long video recordings for non-invasive monitoring in NICU was described. The proposed method has been evaluated on a large amount of data and specially designed for real-life conditions. Automatic extraction of the dynamics of motion organisation in preterm newborns has been performed and observations are consistent with the literature. We also proposed a FMA estimation, based on a compact and interpretable model, to serve as a clinical decision support tool. Although its relevance has yet to be verified in specific clinical applications, this work is an important step towards the integration of non-invasive camera-based

monitoring methods in NICU. Indeed, contrary to previous works, this is the first time that a non-invasive, robust and generalized monitoring tool of the maturation of preterm newborns has been designed, compatible with various care environments of NICU.

Availability of data and materials

In accordance with French regulation, data will be shared through the health-data-hub (https://www.health-data-hub.fr/).

Acknowledgments

Results incorporated in this publication received funding from the European Unions Horizon 2020 research and innovation program under grant agreement no. 689260 (Digi-NewB project). We would like to thank the clinicians, nurses and technicians of the hospitals hosting this work for their motivation and professionalism in carrying out the acquisitions.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] C. A. Lovejoy, V. Buch, and M. Maruthappu, "Artificial intelligence in the intensive care unit," pp. 1–3, 2019.

[2] "Digi-NewB - GCS HUGO - CHU - monitoring system," http://www.digi-newb.eu, accessed 14 April 2020.

[3] S. Cabon, F. Porée, A. Simon, O. Rosec, P. Pladys, and G. Carrault, "Video and audio processing in paediatrics: a review," *Physiological Measurement*, vol. 40, no. 2, p. 02TR02, 2019.

[4] H. F. Prechtl, C. Einspieler, G. Cioni, A. F. Bos, F. Ferrari, and D. Sontheimer, "An early marker for neurological deficits after perinatal brain lesions," *The Lancet*, vol. 349, no. 9062, pp. 1361–1363, 1997.

[5] S. Orlandi, K. Raghuram, C. R. Smith, D. Mansueto, P. Church, V. Shah, M. Luther, and T. Chau, "Detection of atypical and typical infant movements using computer-based video analysis," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2018, pp. 3598–3601.

[6] L. Adde, A. Brown, C. Van den Broeck, K. DeCoen, B. H. Eriksen, T. Fjørtoft, D. Groos, E. A. F. Ihlen, S. Osland, A. Pascal, *et al.*, "In-motion-app for remote general movement assessment: a multi-site observational study," *BMJ open*, vol. 11, no. 3, p. e042147, 2021.

[7] M. Pediaditis, M. Tsiknakis, and N. Leitgeb, "Vision-based motion detection, analysis and recognition of epileptic seizures-A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 3, pp. 1133–1148, 2012.

[8] K. Cuppens, L. Lagae, B. Ceulemans, S. V. Huffel, and B. Vanrumste, "Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy," *Medical & Biological Engineering & Computing*, vol. 48, no. 9, pp. 923–931, sep 2010.

[9] N. B. Karayiannis, G. Tao, J. D. Frost, M. S. Wise, R. A. Hrachovy, and E. M. Mizrahi, "Automated detection of videotaped neonatal seizures based on motion segmentation methods," *Clinical Neurophysiology*, vol. 117, no. 7, pp. 1585–1594, jul 2006.

[10] P. W. Fuller, W. H. Wenner, and S. Blackburn, "Comparison between time-lapse video recordings of behavior and polygraphic state determinations in premature infants," *Psychophysiology*, vol. 15, no. 6, pp. 594–598, 1978.

[11] S. Cabon, F. Porée, A. Simon, B. Met-Montot, P. Pladys, O. Rosec, N. Nardi, and G. Carrault, "Audio- and video-based estimation of the sleep stages of newborns in neonatal intensive care unit," *Biomedical Signal Processing and Control*, vol. 52, pp. 362–370, 2019.

[12] P. R. Berge, L. Adde, G. Espinosa, and Stavdahl, "ENIGMA - Enhanced interactive general movement assessment," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2664–2672, 2008.

[13] D. Karch, K. S. Kim, K. Wochner, J. Pietz, H. Dickhaus, and H. Philippi, "Quantification of the segmental kinematics of spontaneous infant movements," *Journal of Biomechanics*, vol. 41, no. 13, pp. 2860–2867, sep 2008.

[14] M. Donati, F. Cecchi, F. Bonaccorso, M. Branciforte, P. Dario, and N. Vitiello, "A modular sensorized mat for monitoring infant posture." *Sensors (Basel, Switzerland)*, vol. 14, no. 1, pp. 510–531, 2013.

[15] S. Cabon, F. Porée, A. Simon, M. Ugolin, O. Rosec, G. Carrault, and P. Pladys, "Motion estimation and characterization in premature newborns using long duration video recordings," *IRBM*, vol. 38, no. 4, pp. 207–213, 2017.

[16] E. A. Ihlen, R. Støen, L. Boswell, R.-A. de Regnier, T. Fjørtoft, D. Gaebler-Spira, C. Labori, M. C. Loennecken, M. E. Msall, U. I. Möinichen, *et al.*, "Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study," *Journal of clinical medicine*, vol. 9, no. 1, p. 5, 2020.

[17] A. J. Spittle, N. C. Brown, L. W. Doyle, R. N. Boyd, R. W. Hunt, M. Bear, and T. E. Inder, "Quality of general movements is related to white matter pathology in very preterm infants." *Pediatrics*, vol. 121, no. 5, pp. e1184–9, may 2008.

[18] S. Cabon, F. Porée, G. Cuffel, O. Rosec, F. Geslin, P. Pladys, A. Simon, and G. Carrault, "Voxyvi: A system for long-term audio and video acquisitions in neonatal intensive care units," *Early Human Development*, p. 105303, 2021.

[19] M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, and L. Tarassenko, "Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit," *npj Digital Medicine*, vol. 2, no. 1, pp. 1–18, 2019.

[20] R. Weber, S. Cabon, A. Simon, F. Porée, and G. Carrault, "Preterm newborn presence detection in incubator and open bed using deep transfer learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1419–1428, 2021.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] L. Cailleau, R. Weber, S. Cabon, C. Flamant, J.-M. Roué, G. Favrais, G. Gascoin, A. Thollot, F. Porée, and P. Pladys, "Quiet sleep organization of very preterm infants is correlated with postnatal maturation," *Frontiers in Pediatrics*, vol. 8, p. 613, 2020.

[26] X. Ning, I. W. Selesnick, and L. Duval, "Chromatogram baseline estimation and denoising using sparsity (BEADS)," *Chemometrics and Intelligent Laboratory Systems*, vol. 139, pp. 156–167, 2014.

[27] G. M. K. Ntonfo, G. Ferrari, R. Raheli, and F. Pisani, "Low-complexity image processing for real-time detection of neonatal clonic seizures," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 375–382, 2012.

[28] H. Rahmati, R. Dragon, O. M. Aamo, L. Adde, y. Stavdahl, and L. Van Gool, "Weakly supervised motion segmentation with particle matching," *Computer Vision and Image Understanding*, vol. 140, pp. 30–42, 2015.

[29] X. Navarro, F. Porée, M. Kuchenbuch, M. Chavez, A. Beuchée, and G. Carrault, "Multi-feature classifiers for burst detection in single EEG channels from preterm infants," *Journal of Neural Engineering*, vol. 14, no. 4, p. 046015, 2017.

[30] M. Särkelä, S. Mustola, T. Seppänen, M. Koskinen, P. Lepola, K. Suominen, T. Juvonen, H. Tolvanen-Laakso, and V. Jäntti, "Automatic analysis and monitoring of burst suppression in anesthesia," *Journal of Clinical Monitoring and Computing*, vol. 17, no. 2, pp. 125–134, 2002.

[31] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[32] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[33] D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Social psychological and personality science*, vol. 8, no. 4, pp. 355–362, 2017.

[34] R. M. Sakia, "The box-cox transformation technique: a review," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 41, no. 2, pp. 169–178, 1992.

[35] L. Adde, J. L. Helbostad, A. R. Jensenius, G. Taraldsen, K. H. Grunewaldt, and R. StØen, "Early prediction of cerebral palsy by computer-based video analysis of general movements: A feasibility study," *Developmental Medicine and Child Neurology*, vol. 52, no. 8, pp. 773–778, 2010.

[36] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review," *Applied Sciences*, vol. 11, no. 11, p. 5088, 2021.

[37] I. Zuzarte, A. H. Gee, D. Sternad, and D. Paydarfar, "Automated movement detection reveals features of maturation in preterm infants," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*. IEEE, 2020, pp. 600–603.

[38] A. Georgoulas, L. Jones, M. P. Laudiano-Dray, J. Meek, L. Fabrizi, and K. Whitehead, "Sleep–wake regulation in preterm and term infants," *Sleep*, vol. 44, no. 1, p. zsaa148, 2021.

[39] I. Doroniewicz, D. J. Ledwoń, A. Affanasowicz, K. Kieszczyńska, D. Latos, M. Matyja, A. W. Mitas, and A. Myśliwiec, "Writhing movement detection in newborns on the second and third day of life using pose-based feature machine learning classification," *Sensors*, vol. 20, no. 21, p. 5986, 2020.

[40] S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, K. McCormick, A. Zisserman, and L. Tarassenko, "Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning," *Physiological Measurement*, vol. 40, no. 11, p. 115001, 2019.

[41] S. Moccia, L. Migliorelli, R. Pietrini, and E. Frontoni, "Preterm infants' limb-pose estimation from depth images using convolutional neural networks," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE, 2019, pp. 1–7.

[42] C. Einspieler, H. F. Prechtl, F. Ferrari, G. Cioni, and A. F. Bos, "The qualitative assessment of general movements in preterm, term and young infantsreview of the methodology," *Early human development*, vol. 50, no. 1, pp. 47–60, 1997.

[43] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2019.

[44] L. Curzi-Dascalova, "Physiological correlates of sleep development in premature and full-term neonates," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 22, no. 2, pp. 151–166, 1992.

[45] R. Joshi, D. Kommers, L. Oosterwijk, L. Feijs, C. van Pul, and P. Andriessen, "Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics and ECG-derived estimates of infant motion," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 3, pp. 681–692, 2020.

[46] J. Fawke, "Neurological outcomes following preterm birth," in *Seminars in fetal and neonatal medicine*, vol. 12, no. 5. Elsevier, 2007, pp. 374–382.

[47] C. S. Leon, S. Cabon, H. Patural, G. Gascoin, C. Flamant, J.-M. Roue, G. Favrais, A. Beuchee, P. Pladys, and G. Carrault, "Evaluation of maturation in preterm infants through an ensemble machine learning algorithm using physiological signals," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[48] S. Cabon, B. Met-Montot, F. Porée, O. Rosec, A. Simon, and G. Carrault, "Automatic extraction of spontaneous cries of preterm newborns in neonatal intensive care units," in *2020 28th European Signal Processing Conference*. IEEE, 2021, pp. 1200–1204.

[49] B. Met-Montot, S. Cabon, G. Carrault, and F. Porée, "Spectrogram-based fundamental frequency tracking of spontaneous cries in preterm newborns," in *2020 28th European Signal Processing Conference*. IEEE, 2021, pp. 1185–1189.