Continuous prediction and clinical alarm management of late-onset sepsis in preterm infants using vital signs from a patient monitor

Citation for published version (APA):

Document license:
CC BY

DOI:
10.1016/j.cmpb.2024.108335

Document status and date:
Published: 01/10/2024

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 30. Jul. 2024
Continuous prediction and clinical alarm management of late-onset sepsis in preterm infants using vital signs from a patient monitor

Meicheng Yang\textsuperscript{a,b}, Zheng Peng\textsuperscript{b,c}, Carola van Pul\textsuperscript{b,c,d}, Peter Andriessen\textsuperscript{d,e}, Kejun Dong\textsuperscript{f}, Demi Silvertand\textsuperscript{e}, Jianqing Li\textsuperscript{a,g,1}, Chengyu Liu\textsuperscript{a,1}, Xi Long\textsuperscript{b,*}, 1

\textsuperscript{a} State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing, China
\textsuperscript{b} Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
\textsuperscript{c} Department of Clinical Physics, Máxima Medical Centre, Veldhoven, the Netherlands
\textsuperscript{d} Department of Physics, Eindhoven University of Technology, Eindhoven, the Netherlands
\textsuperscript{e} Department of Pediatrics, Máxima Medical Centre, Veldhoven, the Netherlands
\textsuperscript{f} Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, United States of America
\textsuperscript{g} School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China

ARTICLE INFO

Keywords:
Late-onset sepsis
Preterm infants
Artificial intelligence
Predictive modeling
Vital signs
Alarm policy

ABSTRACT

Background and Objective: Continuous prediction of late-onset sepsis (LOS) could be helpful for improving clinical outcomes in neonatal intensive care units (NICU). This study aimed to develop an artificial intelligence (AI) model for assisting the bedside clinicians in successfully identifying infants at risk for LOS using non-invasive vital signs monitoring.

Methods: In a retrospective study from the NICU of the Máxima Medical Center in Veldhoven, the Netherlands, a total of 492 preterm infants less than 32 weeks gestation were included between July 2016 and December 2018. Data on heart rate (HR), respiratory rate (RR), and oxygen saturation (SpO\textsubscript{2}) at 1 Hz were extracted from the patient monitor. We developed multiple AI models using 102 extracted features or raw time series to provide hourly LOS risk prediction. Shapley values were used to explain the model. For the best performing model, the effect of different vital signs and also the input type of signals on model performance was tested. To further assess the performance of applying the best performing model in a real-world clinical setting, we performed a simulation using four different alarm policies on continuous real-time predictions starting from three days after birth.

Results: A total of 51 LOS patients and 68 controls were finally included according to the patient inclusion and exclusion criteria. When tested by seven-fold cross-validations, the mean (standard deviation) area under the receiver operating characteristic curve (AUC) six hours before CRASH was 0.875 (0.072) for the best performing model, compared to the other six models with AUC ranging from 0.782 (0.089) to 0.846 (0.083). The best performing model performed only slightly worse than the model learning from raw physiological waveforms (0.886 [0.068]), successfully detecting 96.1% of LOS patients before CRASH. When setting the expected alarm window to 24 h and using a multi-threshold alarm policy, the sensitivity metric was 71.6%, while the positive predictive value was 9.9%, resulting in an average of 1.15 alarms per day per patient.

Conclusions: The proposed AI model, which learns from routinely collected vital signs, has the potential to assist clinicians in the early detection of LOS. Combined with interpretability and clinical alarm management, this model could be better translated into medical practice for future clinical implementation.

1. Introduction

Neonatal sepsis refers to a systemic bacterial infection in newborns, is a major contributor to neonatal mortality [1]. Preterm infants are particularly vulnerable to both in early and late-onset forms based on sepsis onset [2]. Early-onset sepsis (EOS) usually refers to sepsis in neonates at or before 72 h of life, which is generally caused by vertical bacterial transmission from the mother during the perinatal period.
Late-onset sepsis (LOS) is defined as sepsis onset after 72 h of life as a result of postnatal environmental exposure to pathogenic bacteria. The prevalence of LOS in premature infants varies from 20% to 38% during their initial 120 days of life, with mortality rates spanning from 13% to 19% [3–5]. Recent evidence suggests that early diagnosis of LOS patients and initiation of antibiotic therapy is effective in improving clinical outcomes [6]. A positive blood culture is the gold standard for diagnosing LOS. However, this test has an inherent delay in reporting results due to a natural growth period up which could result in a late clinical diagnosis [7]. Therefore, in actual clinical practice, if there is a suspicion of sepsis in premature infants based on clinical symptoms, broad-spectrum antibiotics are initiated by clinicians when the baby’s blood is taken, rather than waiting for the blood culture results before taking action. This sepsis suspicion time was denoted as a Cultures, Resuscitation, and Antibiotics Started Here (CRASH) moment [6]. However, physiological changes in the patient may have already commenced before this moment when the clinical signs were subtle [9]. Therefore, developing a pragmatic model to help early detection of LOS before CRASH is critical.

With the improvement of medical information systems, the construction of data-driven artificial intelligence (AI) algorithms to identify LOS risk is growing rapidly [10–12]. Among these, electronic health record (EHR) data were commonly used in developing the models [13, 14]. However, the low recording frequency of EHR may still lead to delays in LOS identification. In separate modules in the EHR also vitals can be stored, typically with a frequency once every minute or even more with longer intervals. This information appears useful for sepsis prediction [15, 16], though changes in heart rate variability (HRV), known to change preceding sepsis [17, 18], can not be detected this way. Another approach is to focus directly on the waveforms and HRV information in the patient monitor, which is shown to be useful for predicting impending sepsis [17, 19–21]. However, these high-frequency data places high demands on fast signal processing algorithms, and the significant differences in signal waveforms between different devices make widespread application difficult. An approach in the middle is using vital signs including heart rate (HR), respiratory rate (RR), and oxygen saturation (SpO2) at 0.5 Hz or 1 Hz for prediction modeling [22–24]. However, the effect of input signal types or input data with different sampling rates on model performance and how this can be used for wider application in the NICU, where patients typically require bedside monitors, has not been investigated. In addition, few have attempted to apply AI models in real-world settings, i.e., to quantify in detail the potential clinical impact of continuous real-time prediction and made clinical alarm management. How to integrate interpretability analysis of models with stratified management of patients in long-term monitoring to better support clinicians decision making and reduce the burden on clinical healthcare staff is important.

In the present study, we aim to develop AI-based models to predict the risk of LOS in preterm infants using multiple vital signs data sampled at 1 Hz derived from patient monitors that can be compared to our previous algorithms based on full physiological waveform information [21]. We also test the effect of different vital signs and also the input type of signals on model performance. In addition, we evaluate our models using four different alarm policies to simulate clinical use and manage clinical alarms [15]. The included model explanation is also expected to make the model more readable for translation into clinical practice.

2. Methods

2.1. Study design and population

The data used in this retrospective study were from the neonatal intensive care unit (NICU) of the Máxima Medical Center in Veldhoven, the Netherlands and has been published before [21]. Briefly, a total of 492 preterm infants less than 32 weeks of gestation were included between July 2016 and December 2018. Infants were excluded if their NICU stay was less than 5 days to avoid insufficient clinical and laboratory follow-up for defining sepsis. We also excluded infants with a sepsis onset less than 72 h after birth (EOS), infants with a large intraventricular hemorrhage greater than or equal to grade III, congenital anomalies, or syndrome, and infants without culture-proven LOS even with clinical suspicion of sepsis. Infants were also excluded from this study if there was no sufficiently long record of continuous vital signs from patient monitors. All protected health information was de-identified. The medical ethics committee of the hospital granted a waiver for this retrospective study, in accordance with Dutch law on medical research involving human subjects.

2.2. Sepsis definition

LOS was initially identified based on clinical signs of a generalized infection in accordance with the Vermont Oxford Criteria, isolation of pathogens from a blood culture obtained after the third day of life, and the initiation of intravenous antibiotic treatment [25]. LOS patients were further confirmed and included if the C-reactive protein level was greater than or equal to 10 mg/L at least once within 5 days of the initial identification of LOS, otherwise excluded from the study [26]. We only included the first LOS episode for further analysis. Preterm infants with no clinical suspicion of sepsis and thus no need for blood cultures were included in the control group.

The CRASH moment was identified as the time of suspected sepsis and AI models were aimed at the early detection of LOS risk before the CRASH. As physiological characteristics could change with maturation in premature infants, we first identified the CRASH moment [8] as an ‘anchor point’ for each LOS patient. We then sought one or more control patients with a gestational age (GA) within 3 days either younger or older than the LOS patient. After that, we computed an ‘equivalent CRASH moment’ for these controls by aligning with the postmenstrual age (PMA) to be close to that of the LOS patients to establish a fair period for further analysis, see our previous research for details [19, 21]. This approach allowed a comparison of vital signs around CRASH without the confounding influence of maturation effects.

2.3. Data collection and preprocessing

Vital signs, including HR obtained from electrocardiography, RR determined from the chest impedance (CI) signal, and SpO2 derived from pulse oximetry, were obtained from standard patient monitors used in daily clinical care (Philips IntelliVue MX 800, Philips, Hamburg, Germany). Numerical data at 1 Hz were stored in a data server (PIIC-IX, Data Warehouse Connect, Philips Medical Systems, Andover, MA). Fig. 1 shows the data collection process.

Extreme outliers of vital signs beyond the 0.1st and 99.9th percentiles were removed during model training. Vital-sign time series were divided into hourly units from the first record three days after birth to produce hourly LOS risk predictions. After that, we further split the hourly data into 5-min length non-overlapping window data. For each 5-min window, if there was one of the missing segments longer than the 20 s within that window, the window was removed. Otherwise, missing data were linearly interpolated, or used the next valid observation to fill the gap at the beginning.

2.4. Feature extraction

Two demographic features of GA and birth weight (BW) were included, serving as indicators of the physiological maturity of preterm infants at birth [21]. For each vital sign, we first calculate the features within the segmented 5-min non-overlapping window. We then aggregate the features for each hour by calculating the mean, minimum (min), maximum (max), and standard deviation (SD) across all 12 windows within that hour. Details of each feature type for each vital sign are
defined in the following section.

(1) **HR-based features:** In each 5-min window, the median (Med), SD, and interdecile range (IDR) of the HR sequence were calculated. The sample asymmetry analysis (SampAsy) was computed to describe the reduced accelerations and/or transient decelerations of HR to neonatal sepsis [27]. Average acceleration response (AAR) and average deceleration response (ADR) were designed to capture the average HR response to the acceleration and deceleration, respectively [28]. The percentage of decelerations (pDec), SD of all HR values contributing to pDec (SDDec) were also computed, aiming at explicitly extracting variations from decelerations [29]. Sample entropy (SampEn) was used to quantify the unpredictability of HR sequence fluctuations [30]. We also calculated three features using visibility graph (VG) analysis, which maps a time series into a graph network according to certain geometric criteria in order to better exploit its dynamics or properties [31]. To formalize the VG method, each data point of the time series is transformed into a node, and the connectivity between nodes is defined with the visibility criterion. We calculated the mean degree in VG (VGMD), a measure of the complexity of the network, and the degree variation, the SD of the node degrees (VGSD). The assortativity coefficient (VGAssort) was also computed, where the network is assortative if the connected nodes have a comparable degree (VGAssort > 0), otherwise is disassortative (VGAssort < 0). In addition, the max cross-correlation between HR and SpO2 sequence (HR-SpO2-CrossC) was measured with a lag time of −30 to +30 s after standardizing each sequence (subtracting the mean and dividing by the SD) [32].

(2) **RR-based features:** To measure the asymmetry and the tailedness of the probability distribution of the RR values, the adjusted Fisher-Pearson standardized moment coefficient of skewness (Skew), and kurtosis (Kurt) were computed [21]. For example, the occurrence of apnea in neonatal sepsis may contribute to a prolonged left tail in the RR distribution, potentially manifesting Skew as a negative value. A positive Kurt value indicates more respiration rate values located around the mean rather than in the tails of the distribution. In addition to the two features of RR-Skew and RR-Kurt, four other features were calculated: RR-Med, RR-IDR, RR-SD, RR-SampEn. The details of their calculation were the same as for the HR features mentioned.
**Table 1.** Features within the segmented 5-min non-overlapping window.

<table>
<thead>
<tr>
<th>Vital signs</th>
<th>Feature</th>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart rate (13)</td>
<td>Median</td>
<td>Med</td>
<td>Median value</td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td></td>
<td>Interdecile range</td>
<td>IDR</td>
<td>Defined as the difference between the 90th and 10th percentile</td>
</tr>
<tr>
<td>Sample asymmetry</td>
<td>SampAsy</td>
<td></td>
<td>Describe changes in the shape of the histogram of HR that are caused by</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>reduced accelerations and/or transient decelerations</td>
</tr>
<tr>
<td>Average accelerations</td>
<td>AAR</td>
<td></td>
<td>Capture the average variable response to the acceleration</td>
</tr>
<tr>
<td>response</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average deceleration</td>
<td>ADR</td>
<td></td>
<td>Capture the average variable response to the deceleration</td>
</tr>
<tr>
<td>response</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of the</td>
<td>pDec</td>
<td></td>
<td>Extract variations in time series arising from decelerations</td>
</tr>
<tr>
<td>decelerations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation of</td>
<td>SDDec</td>
<td></td>
<td>Defined as the SD of variable values that contribute to pDec</td>
</tr>
<tr>
<td>deceleration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample entropy</td>
<td>SampEn</td>
<td></td>
<td>Quantify the unpredictability of sequence fluctuations</td>
</tr>
<tr>
<td>Mean degree in visibility</td>
<td>VGMD</td>
<td></td>
<td>The mean degree of a node, where the degree is the number of connections</td>
</tr>
<tr>
<td>graph</td>
<td></td>
<td></td>
<td>(or edges) the node has in the visibility graph analysis</td>
</tr>
<tr>
<td>Standard deviation of</td>
<td>VGSD</td>
<td></td>
<td>Describe degree variation</td>
</tr>
<tr>
<td>node degrees in</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>visibility graph</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assortativity in visibility</td>
<td>VGAssort</td>
<td></td>
<td>A correlation coefficient between the degrees of the nodes on opposite</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ends of an edge</td>
</tr>
<tr>
<td>Cross-correlation</td>
<td>CrossC</td>
<td></td>
<td>The maximum cross-correlation value between two vital signs with a lag time</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>of -30 s</td>
</tr>
<tr>
<td>Respiratory rate</td>
<td>Median</td>
<td>Med</td>
<td>Median value</td>
</tr>
<tr>
<td>(6) and oxygen</td>
<td>Standard deviation</td>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>saturation (6)</td>
<td>Interdecile range</td>
<td>IDR</td>
<td>Defined as the difference between the 90th and 10th percentile</td>
</tr>
<tr>
<td></td>
<td>Sample entropy</td>
<td>SampEn</td>
<td>Quantify the unpredictability of sequence fluctuations</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>Skew</td>
<td>Measure the distortion of symmetrical distribution</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>Kurt</td>
<td>Measure the tailedness of a distribution</td>
</tr>
<tr>
<td>Demographics</td>
<td>Gestational age</td>
<td>GA</td>
<td>Gestational age</td>
</tr>
<tr>
<td></td>
<td>Birth weight</td>
<td>BW</td>
<td>Birth weight</td>
</tr>
</tbody>
</table>

2.5. Model development and explanation

All models were developed using collected multiple vital signs data to predict the risk of LOS in preterm infants (Fig. 1). Four common machine learning algorithms used for sepsis prediction [11], but with very different principles, were constructed for comparative purposes, including, logistic regression (LR), k-nearest neighbors (KNN), support vector machine (SVM), and extreme gradient boosting (XGB). Besides, an LR model only using SD, SampAsy, and SampEn (calculated from HR) was also developed for comparison, as it was the core of the heart rate characteristics (HRC) index monitor (HeRO monitor), which was the first commercially available bedside monitoring tool for neonatal sepsis [17]. We referred to this LR model as the HRC-like model in the following analysis. The above algorithms were all feature-based models, which require feature engineering. We also developed two commonly used deep learning models for time series prediction tasks for comparison, including, an attention-based long short-term memory neural network (Att-LSTM) [33], and a residual convolutional neural network with feature map attention (Att-ResNet) that has been used in our previous research [34]. These two deep learning models were used directly for modeling 1-hour time series data sampled at 1 Hz (length of 3600). Unlike missing value imputation methods in feature-based machine learning, a strategy involving both forward filling and backward filling was utilized. In addition, a 0/1 indicator sequence is created to distinguish between imputed values and actual sampled values.

Seven-fold cross-validation was used to train and evaluate all the seven prediction models. The patients were divided into seven subsets or folds, meaning that the models were trained and tested seven times, each time using a different subset/fold as the hidden test set. We only used the data from the 24 h before CRASH in each patient for training. In addition, further seven-fold cross-validation was performed within each training set to optimize the positive window length (number of hours before CRASH in LOS patients as positive samples) and hyperparameters of the models. We employed the Bayesian optimization method [35] to optimize the model hyperparameters (Supplementary Table 1). The final prediction was made using the ensemble approach by averaging the prediction probability of the seven models generated during the hyperparameter optimization. To address the issue of class imbalance, class weights were used to adjust the importance of each class during training. When testing on each hidden test set, to maintaining consistency with the strategy used in previous research for assessing model performance [20,21], we used the predictions during the 6 h prior to CRASH to determine the best performing model among all the developed seven models for the following analysis. Metric used was the area under the receiver operating characteristic curve (AUC) and reported as mean (SD) due to the seven-fold cross-validation.

To evaluate the predictive value of different vital signs or their combinations in the best performing model, we considered six feature sets for comparison: HR features, RR features, SpO2 features, the combination of HR and RR features, the combination of HR and SpO2 features, the combination of RR and SpO2 features. The demographic features of GA and BW were included in all the feature sets. We also calculated hourly AUC values for different feature sets comprising different combinations of vital signs. This process involved collecting predictions from each test fold for all patients at each hour of the 24-hour prior to CRASH. This approach allowed us to analyze prediction performance in terms of ‘time to CRASH’ [21].

In addition, after obtaining the best performing model, to test the effect of different signal types and different sampling rate of vital signs on the predictive performance, we first retrained and tested our previously published model [21], an XGB model constructed based on the...
electrocardiogram (ECG) (250 Hz) and CI (62.5 Hz) waveforms, using the same patients in this study. In addition, we down-sampled the second-by-second vital signs data to minute-by-minute and hour-by-hour, respectively, and then retrained and tested. For the minute-by-minute data, there were 60 points in an hourly sequence, and the resulting 27 features within each hour also followed the features in Table 1. For the hour-by-hour data, there was only one point in per hour, consisting of five features: HR, \(\text{SpO}_2\), RR, GA, and BW.

Model was explained using the SHapley Additive exPlanations (SHAP) [36] method. SHAP values are from the game theory literature, providing a fair way to distribute the ‘credit’ among the features by considering all possible combinations of features and measure their impact on the prediction. We use the SHAP method to explain the model from two perspectives: individual interpretability and global understanding. Individual interpretability means using SHAP values to assign an impact score to each feature for every hourly prediction for a single patient. This helps to explain how the model outputs risk probabilities and could help to highlight relevant risk factors. Global understanding can show the overall importance of a specific feature to the model by averaging the impact scores across all patients.

2.6. Evaluation of continuous prediction

After getting the best performing model, we simulated its use in a real clinical setting, i.e., making continuous real-time (hourly) predictions from 72 h after birth (Fig. 2A). We limited the prediction duration to 14 days for some patients may have very long ICU stays. For LOS patients, if the duration from birth+72 h to the occurrence of CRASH exceeds 14 days, only the data from the 14 days preceding CRASH is considered; otherwise, data between birth+72 h and the time of CRASH is utilized. For control patients, if the time from birth+72 h to the timestamp of the last vital sign record exceeds 14 days, only the data from the first 14 days is considered; otherwise, data from birth+72 h to the time of the last vital sign record is used. We then presented the predicted probability per hour by calculating the average of the predicted probability in test sets for all patients at each hour before CRASH for LOS patients and the averaged predicted probability per hour after birth+72 h for controls.

A clinician would be alarmed if an hourly probability prediction exceeded a certain threshold. We set an expected alarm window (e.g., [CRASH-24, CRASH]), i.e., any alarm within this window was considered a true positive. Any alarm within the expected alarm window was considered a true positive (TP) (Fig. 1). Alarms occurring outside of the time window were treated as false positives (FP), while the absence of alarms within the window was considered as a false negative (FN). However, the continuous hourly prediction would result in a large number of false alarms and lead to alarm fatigue [37]. To address this, a traditional silencing policy was applied, which means the model would be silenced for a period after an alarm was triggered [38,39]. The silencing period was set to 8 h following a recent research [16].

A probability prediction threshold is important in a clinical setting. We performed a simulation to evaluate the effect of different probability prediction thresholds and different alarm policies in a clinical setting. In this simulation, we defined three different probability thresholds (low, medium, high) based on the alarm rate in hours without the silencing policy. Following a recent research of van den Berg et al. [16], these thresholds correspond to raise three alarms per day, one alarm per day, or one alarm per week per patient in the test patients, respectively. Based on these thresholds, we then tested three different alarm policies: low-threshold alarm with silencing policy, medium-threshold alarm policy, high-threshold alarm policy.
with silencing policy, high-threshold alarm with silencing policy. In addition, we also used the multi-threshold alarm policy, based on a recent research for testing [16], as shown in Fig. 2B. In this scenario, an alarm would also be triggered during the 8-hour refractory period if it corresponded to a higher risk level. Finally, a total of four alarm policies were tested.

We set four different expected alarm windows (the time period in which alarms should be triggered preceding the CRASH) for evaluation: [CRASH-12, CRASH], [CRASH-24, CRASH], [CRASH-36, CRASH], and [CRASH-48, CRASH], for clinicians may have different requirements for alarms under different alarm policies in the continuous prediction, we could report the performance metrics of sensitivity and positive predictive value (PPV) (alarm-wise). Average alarms per day per patient for LOS, control, and all patients were also reported.

In addition, in order to know how many patients the model could successfully predict or detect before CRASH, we treated the patients as positive cases (LOS patients) if there was at least one prediction before CRASH whose probability was higher than the threshold. Therefore, the LOS/non-LOS predictions for all patients could be identified patient-wise, so that the metrics of sensitivity, specificity, and PPV can be reported (patient-wise).

3. Results

3.1. Study population

A total of 51 LOS patients and 68 control patients were finally included, as shown in Fig. 3, Table 2 describes the baseline characteristics between the included LOS and control patients. Hospital mortality was higher in patients diagnosed with LOS than in those without (13.7 % versus 2.9 %). Compared with controls, LOS patients had a longer median (IQR) GA (27.86 [26.71–29.21] versus 29.43 [28.71–30.71] weeks) and lower BW (1030 [850–1265] vs 1268 [1075–1409] g). LOS patients were more likely to have a patent ductus arteriosus (ibuprofen) and used surfactant replacement therapy. In-hospital mortality was higher in patients diagnosed with LOS than in those without (13.7 % versus 2.9 %). Compared with controls, LOS patients had a longer mean (±SD) hospital stay (2.9 ± 0.9 days) than controls, showing a characteristic of an initial increase followed by a subsequent decrease. On the other hand, HR-pDec and HR-VGMD showed little fluctuation, maintaining a relatively stable condition during this period. For LOS patients, regarding the HR-based features, the HR-Med exhibited significant fluctuations, yet consistently remains greater than controls, prior to CRASH. The HR-IDM displayed a trend of an initial increase before CRASH, followed by a descent, indicating faster accelerations than decelerations. In the case of HR, the model including all vital signs showed the best performance with an AUC of 0.875 (0.072). For SpO2-based features, SpO2-Med remained consistently lower than controls, but SpO2-SD remained consistently lower than controls, but SpO2-Med was higher than controls, showing a characteristic of an initial increase followed by a subsequent decrease.

3.2. Model performance and explanation

Fig. 5A shows the prediction performance of the seven developed models six hours prior to LOS onset in the form of receiver operating characteristic (ROC) curves. The developed XGB model performed best with an AUC of 0.875 (0.072), compared to three machine learning models ranging from 0.799 (0.079) to 0.846 (0.083), two deep learning models of 0.782 (0.089) and 0.801 (0.098), and the HRC-like model of 0.827 (0.109). The results in Fig. 5B indicated that the XGB model using a combination of vital signs outperforms models built on each vital sign individually. For models built from SpO2 and RR alone, the AUCs were 0.749 (0.109) and 0.767 (0.099) respectively. However, when combined, the AUC increased to 0.786 (0.097). In the case of HR, the model constructed on HR alone gave an AUC of 0.846 (0.085). When combined with RR and SpO2 separately, the AUC improved slightly to 0.853 (0.088) and 0.864 (0.080) respectively. The model including all vital signs showed the best performance with an AUC of 0.875 (0.072). Fig. 5C shows the time-varying AUCs for models constructed individually from each vital sign and the model constructed by combining all vital signs in the 24 h prior to the CRASH event. It is shown that the

### Table 2

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Late-onset sepsis (n = 51)</th>
<th>Control (n = 68)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gestational age (weeks)</td>
<td>27.86 (26.71–29.21)</td>
<td>29.43 (28.71–30.71)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Female gender</td>
<td>22 (43.1)</td>
<td>39 (57.4)</td>
<td>0.177</td>
</tr>
<tr>
<td>Birth weight (g)</td>
<td>1030 (850–1265)</td>
<td>1268 (1075–1409)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Singletons</td>
<td>36 (70.6)</td>
<td>46 (67.5)</td>
<td>0.886</td>
</tr>
<tr>
<td>Appar score m min</td>
<td>7 (6–8)</td>
<td>8 (7–8)</td>
<td>0.272</td>
</tr>
<tr>
<td>Umbilical artery PH</td>
<td>7.29 (7.25–7.34)</td>
<td>7.31 (7.26–7.33)</td>
<td>0.574</td>
</tr>
<tr>
<td>Intrapartum maternal antibiotic therapy</td>
<td>15 (29.4)</td>
<td>21 (30.9)</td>
<td>0.977</td>
</tr>
<tr>
<td>Patent ductus arteriosus (ibuprofen)</td>
<td>24 (47.1)</td>
<td>9 (13.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Patients with a central line 48 h before CRASH</td>
<td>23 (45.1)</td>
<td>6 (8.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CRASH in postnatal age (days)</td>
<td>6.92 (5.04–10.18)</td>
<td>5.17 (3.61–6.77)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Postmenstrual age (weeks)</td>
<td>29.02</td>
<td>30.13</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>7 (13.7)</td>
<td>2 (3.0)</td>
<td>0.064</td>
</tr>
</tbody>
</table>

* Continuous variables are expressed as median with interquartile range, and two groups are compared using the Wilcoxon rank-sum test. Categorical variables are presented as numbers and percentages, and two groups are compared using Chi-square tests.
model integrating all vital signs almost consistently outperforms the models built from individual vital signs alone from CRASH-13 h to the CRASH event. In addition, each model reached its maximum AUC at the CRASH-1 h time point.

Compared to the model built on ECG and CI waveform data, the XGB model trained on second-by-second vital signs data showed only a slightly lower AUC performance of 0.875 (0.072) versus 0.886 (0.068), as shown in Table 3. However, as the resolution of the input vital signs

Fig. 4. Feature trends over time for LOS patients and controls. Points represent the mean feature values over all patients. The vertical black dashed line represents the moment of the CRASH event.
data decreased, the AUC of the model gradually decreased, with the corresponding AUC performance for minute-by-minute and hour-by-hour data being 0.825 (0.078) and 0.687 (0.116), respectively.

The impact of features on the XGB model across patients in Fig. 5D shows that higher HR-SpO\textsubscript{2}-CrossC\textsubscript{mean}, lower GA, higher HR-SampA\textsubscript{sy}\textsubscript{mean}, lower HR-pDec\textsubscript{mean} and lower HR-SampEn\textsubscript{mean} were associated with a higher risk of LOS. The ranking of feature importance does not change very significantly over time with the sliding 4-hour window 24 h before CRASH, as shown in Supplementary Figure 1.

### 3.3. Continuous prediction and clinical alarm management

Next, for the XGB model that performed best using multiple vital signs data, we applied it in long-term real-time prediction all timepoints up to 14 days. The median (IQR) predicted duration from birth + 72 h was 3.9 (2–7) days (equivalent to 94 [48-168] hours) for LOS patients and 10.6 (6.8–14) days (equivalent to 254 [162–335] hours) for controls. From Fig. 6, it can be seen that for the LOS patients, the average predicted probability of the model starts to increase significantly around the time of CRASH-24 h. For the controls, the average predicted probability fluctuates minimally and remains lower than for the LOS population at all time points. In addition, the frequency of alarms (without the 8 h silence period) generated in the LOS patients is higher than in the controls, and the number of medium and high threshold alarms is greater.

A higher threshold would result in a higher PPV but would result in
lower sensitivity (Table 4). This means that a higher threshold could produce fewer false alarms, reducing the average number of alarms per person per day. However, it would produce fewer true alarms during the expected alarm period, and many LOS patients who should have triggered an alarm during their ICU stay would be missed. The multi-threshold alarm policy was established based on the low-threshold, the difference was that it would trigger an alarm if a higher threshold was exceeded during the traditional 8-hour silence period. In this case, it could help improve both the PPV and sensitivity in terms of alarm-wise while ensuring that LOS is not missed on a patient-wise basis. This is explained by the possibility of having another alarm during the refractory period of 8 h if the risk probability is higher than that of the previous alarm risk which is not included in the expected alarm window. When expected the alarms in the 24-hour period prior to CRASH, the multi-threshold alarm policy achieved a sensitivity of 71.6 % and a PPV of 9.9 % in alarm-wise. In addition, performance in the patient statistics (patient-wise) showed that the model identified the majority of the LOS patients (sensitivity = 96.1 %). However, it incorrectly identified many non-LOS patients as LOS (specificity = 19.1 %). The average number of alarms per person per day was much higher in LOS patients than in controls (2.02 versus 0.83).

We have illustrated two individual examples of real-time prediction starting from three days after birth for a patient who was eventually diagnosed as LOS and a patient who was diagnosed as non-LOS (Fig. 7).

Table 4
Alarm analysis applied XGB in a real clinical setting when making real-time predictions.

<table>
<thead>
<tr>
<th>Evaluation perspective</th>
<th>Expected alarm window</th>
<th>Metrics</th>
<th>Low-threshold</th>
<th>Medium-threshold</th>
<th>High-threshold</th>
<th>Multi-threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alarm-wise</td>
<td>[CRASH-12, CRASH]</td>
<td>Sensitivity</td>
<td>55.0</td>
<td>41.0</td>
<td>16.0</td>
<td>72.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPV</td>
<td>6.2</td>
<td>10.4</td>
<td>13.5</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>[CRASH-24, CRASH]</td>
<td>Sensitivity</td>
<td>54.7</td>
<td>39.9</td>
<td>14.9</td>
<td>71.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPV</td>
<td>9.2</td>
<td>15.9</td>
<td>19.2</td>
<td>9.9</td>
</tr>
<tr>
<td></td>
<td>[CRASH-36, CRASH]</td>
<td>Sensitivity</td>
<td>54.6</td>
<td>39.2</td>
<td>13.9</td>
<td>69.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPV</td>
<td>12.3</td>
<td>19.9</td>
<td>23.1</td>
<td>13.7</td>
</tr>
<tr>
<td></td>
<td>[CRASH-48, CRASH]</td>
<td>Sensitivity</td>
<td>47.3</td>
<td>32.4</td>
<td>10.5</td>
<td>62.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPV</td>
<td>15.3</td>
<td>23.9</td>
<td>26.9</td>
<td>16.8</td>
</tr>
<tr>
<td>Patient-wise</td>
<td>–</td>
<td>Sensitivity</td>
<td>96.1</td>
<td>86.3</td>
<td>56.9</td>
<td>96.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>19.1</td>
<td>52.9</td>
<td>80.9</td>
<td>19.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PPV</td>
<td>47.1</td>
<td>57.9</td>
<td>69.0</td>
<td>47.1</td>
</tr>
<tr>
<td>LOS</td>
<td>Average number of alarms per patient per day</td>
<td>1.53</td>
<td>0.81</td>
<td>0.24</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>0.79</td>
<td>0.23</td>
<td>0.04</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>0.92</td>
<td>0.39</td>
<td>0.09</td>
<td>1.15</td>
<td></td>
</tr>
</tbody>
</table>

a Alarm-wise. We established an alarm window and classified any alarm occurring within this window as a true positive. Alarms outside this window were classified as false positives, and the absence of alarms within this window was classified as false negatives.

b Patient-wise. Patients would be referred to as positive cases (LOS patients) if at least one prediction during their stay after birth+72 h whose probability was higher than the threshold and negative cases if not.
At the top of Fig. 7, the real-time prediction results are shown, along with instances of actual alarms in the multi-threshold alarm policy. There was only one alarm for patient B, but eight alarms for patient A. Although patient A initially had a few sporadic false alarms, these can be considered as initial indications. What is significant is the occurrence of continuous, higher level alarms within 5 h before the CRASH, which may lead to increased clinician attention to this patient.

The middle of the figure shows one selected hour of vital sign data at 1 Hz. For patient A, many transient decelerations were observed in the HR time series. Oxygen saturation showed instability and remained predominantly at relatively low levels. Respiratory fluctuations were also pronounced, with a relatively high frequency. In patient B, significant fluctuations in the three vital signs were much less pronounced. At the bottom of the figure are the risk factors and their corresponding values, which were obtained by the interpretable SHAP ranking.

4. Discussion

The present study developed seven AI models using multiple vital signs derived from the patient monitor to predict LOS risk in preterm infants admitted to the NICU. Experimental results showed that the best performing model for LOS risk prediction 6 h before CRASH onset was a gradient boosting machine, learning from a total of 102 features extracted based on 1 Hz sampled HR, RR, SpO\textsubscript{2} data, and demographics. We also evaluated the effect of different combinations of vital signs and different type of inputs signals on the model. When using the best performing model to provide continuous real-time prediction of LOS risk starting from birth+72 h, we tested four different alarm policies including a multi-threshold alarm policy, trying to reduce false alarms. Combined with the model explanation, this might be better translated into medical practice for future clinical implementation.

Previous studies have demonstrated using heart rate characteristics and HRV from the high-frequency ECG signal can effectively predict impending LOS [17,18,20]. The addition of respiratory signal features and analysis of the continuous motion index of ECG waveforms has been shown to add information to HR monitoring alone, improving model performance from 0.82 to 0.88 [19,21]. We agree that direct analysis and modeling of the raw physiological waveform data can reveal more information, allowing for a finer-grained analysis of the patient’s disease progression status. However, developing these algorithms for practical clinical practice requires close collaboration with medical device manufacturers to be able to run them real-time on the patient monitor, with rapid raw physiological signals processing. Or, when used on a vendor-neutral system for processing, it needs real-time extraction from the patient monitor and the model then needs to be robust to slight variations in the raw signals from different vendors. These issues can lead to difficulties in large-scale deployment and even performance degradation. Analyzing numerical data from vital signs obtained from the monitor, Kausch et al. [23] developed machine learning models using numerical HR and SpO\textsubscript{2} data sampled at 0.5 Hz for predicting LOS within 24 h and validated them in 3 NICUs (AUCs > 0.78). Honoré et al. [22] combined multiple vital signs of HR, RR, and SpO\textsubscript{2}, which were sampled at 1 Hz, to develop models for predicting LOS and achieved an AUC of 0.82 up to 24 h before clinical sepsis suspicion. And van den Berg et al. [16] used LR, generalized additive models, and XGB modeling minute-by-minute sampled HR and SpO\textsubscript{2} data for predicting LOS risk,

![Fig. 7. Illustrated examples of the individual real-time prediction starting from three days after birth for a diagnosed LOS patient (A) and control patient (B).](image-url)
achieving AUCs ranging from 0.72 to 0.79 at the time of sepsis. In the present study, we used three vital signs (HR + RR + SpO₂) sampled at 1 Hz, generating a total of 102 features. Similar to van den Berg, we observed small differences between LR and XGB, where we found that XGB is slightly outperforming LR, which we also observed in previous studies using features from raw signals (ECG + CI) as input of the model [21].

Compared to that model using the raw waveform signals, the AUC of the XGB model using 1 Hz HR + RR + SpO₂ data decreased slightly from 0.886 (0.068) to 0.875 (0.072). Using 1 Hz HR + RR without SpO₂, the AUC was lower at 0.853 (0.080). The raw waveform signals contained more detailed information and the motion features extracted from ECG also contributed to the better performance. However, this present study proves that almost the same accuracy can be achieved by using numerical data, with the advantage of less storage space, higher computational efficiency, and no need to develop specialized signal processing algorithms. This may be more important in terms of practical application, allowing this method to be used in NICUs of other hospital institutions.

Therefore, in the NICU where patients generally require bed-side monitors, modeling the exported 1 Hz HR + RR + SpO₂ vital sign sequences may become an important method for monitoring LOS risk in preterm infants and has broad application prospects.

AI-based predictive models in healthcare are required to be readable, transparent, and interactive [41,42]. Therefore, when using vital sign sequences for modeling, it is important to interpret these signals and translate them into meaningful data for the clinician at the bedside. Maturation does affect vital signs in preterm infants, but as shown in Fig. 4, many features do not show significant differences between LOS and controls around 24 h before the CRASH. Instead, the gap widens as LOS occurs. This indicates that these features do indeed change as the patient’s condition progresses. Therefore, this shows that incorporating and modeling these dynamic changes in vital signs to predict LOS is effective. Therefore, those features that truly reflect the underlying changes in LOS progression are more important to the model. It is also important to understand why a risk alert occurs and what the physiological mechanisms are [43]. By analyzing the changes of the features extracted over time and model explanations, we found that a higher HR-SampAsy-mean, a lower HR-pDec-mean, and a lower HR-SampEn-mean were associated with higher LOS risk, consistent with previous findings [19,21,27,30,44]. This might be caused by reduced accelerations and/or transient decelerations (inverted spikes) of HR, combined with longer and deeper decelerations, as shown in the left middle of Fig. 7. Entropy falls in the presence of spikes in a record with reduced variability rather than a change in regularity, as has been explained in the Lake et al. research [30]. The transient, repetitive decelerations in HR may be due to vagus nerve firing as part of the cholnergic anti-inflammatory response [45]. In addition, we found that an increased cross-correlation between HR and SpO₂ was associated with a higher risk of LOS, in line with previous research [32,46]. This may be correlated with central apnea events or periodic breathing [47]. Another typical characteristic of neonatal sepsis is persistent tachycardia [48], and we can find that HR-Mean-mean increased before CRASH and was persistently higher than in control. The lower HR-VGMD-mean for LOS means nodes have fewer connections in the visibility graph, indicating a decline in HRV [30]. Depressed HRV indicates autonomic nervous system dysfunction, which may be due to acute or chronic pathology, particularly evident in sepsis [49]. In terms of respiration, LOS patients are observed to have rapid breathing, which may be due to direct pathogenesis in the respiratory system, and perhaps a mechanism to compensate for metabolic acidosis [48]. In addition, an increase in central apnea is also one of the most commonly cited signs of sepsis in preterm infants [46]. We found that a higher RR-SDQ was associated with a higher risk of LOS and was among the top 15 most important features. Hypoxia occurs in neonates with sepsis and may be intermittent or prolonged [48]. In our study, higher SpO₂-SDQ-mean is seen before LOS and in the model this feature contributes in the top-15 feature importance.

In addition, although the ranking of the importance of the top 15 features does not change significantly over time within the sliding 4-hour windows 24 h before the CRASH, it is noteworthy that we observed the importance of the HR-SpO₂-CrossC-mean feature gradually increased (indicated by darker color) as the CRASH approached, while the impact of GA, although still ranked second, slightly decreases (indicated by a lighter color), suggesting that physiological indicators play a more important role in predicting sepsis. More importantly, when using our XGB model in a simulation of clinical setting, we were able to show how risk has changed over time and provide relevant explainable risk factors, as shown in Fig. 7. This made the model readable, which is important for developing a user-friendly interface in daily healthcare practice, enabling users to interact effectively with the system. Therefore, the most important features of our model have the potential to serve as physiomarkers for predicting LOS risk.

The use of a real-time critical illness risk prediction model in clinical practice is capable of capturing instantaneous changes in patient risk. However, it could generate a significant number of false positives due to various noise and disturbance factors, leading to alarm fatigue [37]. Previous studies have used a silencing policy, where the model is silenced for a period after an alarm is triggered, to reduce the frequency of alarms [38,39]. However, if the condition of patients deteriorates rapidly during this period, it is difficult to capture this in a timely manner. Therefore, it is important that higher-level alarms allow clinicians to capture immediate changes in patient risk, similar to the use of regular yellow and red alarms in patient monitoring that ensures that those alarms are considered a different level of alert, allowing for different levels of attention to be given, facilitating stratified management. Nevertheless, due to the characteristics of LOS as a clinical syndrome, predicting LOS remains a challenging task, with many false alarms still occurring. There is still a long way to go for future research, but the patient stratification management strategy based on different levels of alarms validated in this study is crucial and may reduce the workload of clinicians to some extent.

This study has several limitations. First, our study population is a selected group, where only clear LOS positive patients and controls are included, ideal for modeling but meaning that tests need to be performed on the algorithm on a more realistic population. In addition, there is a difference in maturation between LOS patients and controls. It is known that more immature infants have a higher risk of sepsis, so this may influence those characteristics that are known to be maturation-dependent. Future studies should develop the current algorithm on a large matched dataset with no significant difference in maturation between the LOS group and controls. Second, the data set used in this study is relatively small and lacks external validation or validation on a more real NICU population. Future studies will need to collect more data and validate the model in NICUs at other healthcare institutions. Third, the features used in this study are mostly those that have been proven effective in existing literature and no new effective features have been developed. Future research needs to further extract new features from RR and SpO₂ time series and explore the features such as HR-SpO₂-CrossC-mean that can reflect the significant information hidden in the interaction between physiological signals. Fourth, the deep learning models performed relatively poorly. The method for dealing with missing data may be a problem. However, the limited data may contribute more to this problem. Future studies need to develop or test more optimized methods for handling missing values and addressing class imbalance, and collect more data for training to improve the performance of the proposed LOS risk prediction model. Whether or not it can actually improve patient outcomes remains unknown. Further prospective and randomized controlled trials are needed to validate this. Sixth, although we have simulated alarm analysis under different thresholds and tested a multi-threshold alarm policy, no clinical unified standards for choosing the threshold and to assessing which is better. Adjustments still need to be made according to the actual clinical needs. Nevertheless, our proposed approach using multiple vital sign numerical data for raising alarms of LOS risk still offers a new perspective and
solution for clinical practice as a reference, which may be useful for translation into medical devices.

5. Conclusion

In this study, we demonstrated the potential of predicting LOS risk in preterm infants using non-invasive 1 Hz sampled vital signs monitoring rather than raw physiological waveform signals, which may have broader application prospects with the advantage of less storage space and higher computational efficiency. Model explanation allows clinicians to immediately identify risk factors, allowing for transparency in the decision-making process. In addition, the multi-threat alarm policy was shown to have the potential to assist in patient stratification management. Future work could explore the synergistic effects between vital sign sequences to improve the ability to predict sepsis. However, it is more important to design prospective studies to validate the model in a larger population, thereby improving the model to make it more readable, transparent, and interactive, ultimately assisting clinicians in efficiently managing patients.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2024.108335.

References

M. Yang et al.


