REVIEW ARTICLE

# Artificial intelligence and machine learning for real-time late-onset neonatal sepsis prediction using continuous physiologic monitoring: a systematic review and meta-analysis

Layla Abdullah Alajlan[1]*, Aseel M Alsamaani[1], Rola H Alseghair[1], Sarah A Alomar[1], Hind H Almazyad[1], Razan A Aljameeli[1], Abdulrahman Obaid Alharbi[2]

## ABSTRACT

**Background:** Late-onset neonatal sepsis (LOS) causes significant mortality and morbidity in the neonatal intensive care unit (NICU). Early detection is challenging due to non-specific symptoms and low sensitivity of current diagnostic methods. Artificial intelligence (AI) and machine learning (ML) may enable earlier detection through continuous physiologic monitoring. This systematic review evaluates the performance and clinical utility of AI/ML models for early LOS prediction in NICU patients.

**Methods:** Following Preferred Reporting Items for Systematic Reviews and Meta-Analysis 2020 guidelines, PubMed, Google Scholar, and Web of Science were searched for studies using AI/ML to predict LOS from continuous vital signs in the NICUs. Eligible studies reported area-under the curve from receiver operating characteristics curve (AUC-ROC) as the primary outcome, pooled using random-effect models. Subgroup analyses were conducted by model type. Quality was assessed using PROBAST.

**Results:** Ten studies encompassing 8,882 patients were included. The pooled AUC-ROC was 0.84 (95% CI: 0.77-0.92), with high heterogeneity ($I^2$ = 99.2%). Random forest and support vector machine were top performance (AUC = 0.86). Neural network showed lower predictive ability (AUC = 0.78) but no significant heterogeneity ($I^2$ = 0%). Combining multiple vital signs enabled the prediction of sepsis 6 to 48 hours prior to the onset of sepsis, with sensitivity reaching 96%. Multilevel analysis indicated that heterogeneity was driven by intra-study effect size rather than inter-study variance. Half of the studies carried low risk of bias.

**Conclusion:** AI and ML models show promise for early LOS prediction, potentially enabling timely intervention without laboratory data, particularly in resource-limited NICUs; however, substantial heterogeneity warrants cautious interpretation and further validation.

**Keywords:** Late-onset neonatal sepsis, artificial intelligence, machine learning, neonatal intensive care unit, physiologic monitoring.

## Key Messages

- Artificial intelligence and machine learning models using continuous vital sign monitoring predict late-onset neonatal sepsis with good accuracy.
- These models can identify sepsis 6-48 hours before clinical diagnosis, enabling earlier intervention.
- Non-invasive monitoring approaches may improve sepsis surveillance and clinical decision-making in neonatal intensive care units without requiring laboratory data.

**Correspondence to:** Layla Abdullah Alajlan
*College of Medicine, Qassim University, Saudi Arabia
**Email:** Laylaalaajlan@gmail.com
*Full list of author information is available at the end of the article.*
**Received:** 09 October 2026 | **Revised (1):** XX XX XXXX | **Accepted:** 19 October 2026

OPEN ACCESS
This is an open access article distributed in accordance with the Creative Commons Attribution (CC BY 4.0) license: https://creativecommons.org/licenses/by/4.0/) which permits any use, Share — copy and redistribute the material in any medium or format, Adapt — remix, transform, and build upon the material for any purpose, as long as the authors and the original source are properly cited. © The Author(s) 2026.
AMEM

## Introduction

Late-onset neonatal sepsis (LOS) is a bloodstream infection that occurs after the first 72 hours of life in neonates admitted in the neonatal intensive care unit (NICU). LOS remains a leading cause of morbidity and mortality, especially among premature neonates or neonates with low birth weight [1,2]. Unlike early-onset sepsis, which arises from vertical transmission of pathogens from the maternal environment before or during delivery, LOS is predominantly acquired postnatally from the hospital environment [3,4]. The incidence of LOS in hospitalized patients reaches up to 14.2% and increases up to approximately 20% in very low birth weight infants [5]. The global burden of LOS is substantial. In 2021, an estimated 70,411 deaths were attributed to LOS worldwide, reflecting its significant contribution to neonatal mortality [6]. The burden of neonatal sepsis is disproportionately higher in low- and middle-income countries (LMICs), where limited resources and delays in diagnosis contribute to increased incidence and mortality [7].

A critical challenge in the management of LOS is the delay in diagnosis. Early clinical signs are often subtle and nonspecific, making differentiation from other neonatal conditions difficult [8,9]. Blood culture, the current gold standard diagnostic method, has low sensitivity and a prolonged turnaround time, often requiring up to three days for confirmation. This limitation, combined with the lack of universally accepted sepsis definitions, especially in preterm infants, complicates the clinical decision-making and makes the development of a reliable diagnostic method an urgent need [10,11]. As a result of this uncertainty, clinicians frequently initiate empirical broad-spectrum antibiotic therapy, which may lead to unnecessary treatment and contribute to antimicrobial resistance [12]. In addition to diagnostic challenges, LOS imposes a significant economic burden on healthcare systems. Prolonged NICU hospitalization, increased use of invasive monitoring, and extended antibiotic exposure substantially increase healthcare costs, particularly in resource-limited settings [13,14].

In recent years, there has been growing interest in using artificial intelligence (AI) and machine learning (ML) to detect LOS. These approaches utilize continuous monitoring of physiological data, including heart rate variability, respiratory patterns, and oxygen saturation, to identify subtle preclinical changes that precede overt clinical deterioration [15,16]. The high potential of AI and ML offers promising benefits to the LMICs, where the laboratory diagnostics are often costly with long turnaround times [17,18]. The aim of this systematic review is to evaluate the performance and clinical utility of AI and ML models based on continuous vital sign monitoring for early prediction of neonatal sepsis in NICU patients, with particular focus on prediction accuracy, lead time before clinical diagnosis, and applicability in real-world NICU settings.

## Method

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [19]. The review has been registered in the International Prospective Register of Systematic Reviews under the following protocol number (ID: CRD420261338545).

### Search strategy

A comprehensive literature search was carried out from 2015 to March 2026 across PubMed, Google Scholar, and Web of Science. To ensure a comprehensive result, the search employed the following key terms: (neonatal OR newborn) AND (sepsis OR clinical deterioration OR mortality) AND (artificial intelligence OR machine learning) AND ("heart rate" OR "respiratory rate" OR SpO2 OR temperature) AND (continuous OR vital signs OR monitoring) AND (NICU OR "neonatal intensive care"). The search strategy was adapted for each database to account for differences in indexing terms and search syntax.

### *Eligibility criteria*

This review included cohort studies utilizing real NICU data, published between 2015 and 2026 in English. Eligible studies enrolled neonates aged 28 days or younger admitted to NICUs with continuous vital signs monitoring, in which AI/ML prediction models trained on continuous physiologic data from bedside monitors, including heart rate, respiratory rate, oxygen saturation, temperature, blood pressure, and/or perfusion index, were used to predict sepsis onset. Studies were required to report at least one quantitative performance metric such as AUC-ROC, sensitivity, or specificity, and comparison with standard or traditional diagnostic methods was considered. Studies were excluded if all reported models required non vital sign inputs as mandatory features with no vital signs only variant reported, or if they involved animal, simulation, or healthy volunteer populations. Studies lacking continuous vital sign-based prediction outcomes, accessible full text, or English language content were excluded.

### *Study selection*

Two independent reviewers initially screened the retrieved records by title and abstract, in accordance with the predetermined eligibility criteria, using Rayyan (version 1.5.6; Rayyan system inc., Cambridge, MA). Subsequently, full-text screening of the eligible studies was conducted. Any disagreements were resolved through discussion with a third reviewer (Figure 1).

### *Data extraction*

Two authors independently extracted the outcome of interest for our systematic review and meta-analysis: area-under the curve from receiver operating characteristics curve (AUC-ROC). The summary and study characteristics included study ID, study characteristics including country, study type, sample size, patient demographics including birth weight, gestational age, percentage of males, predictive variables: temperature, respiration, heart rate (HR), oxygen saturation, blood pressure, perfusion index, and the purpose of predictive

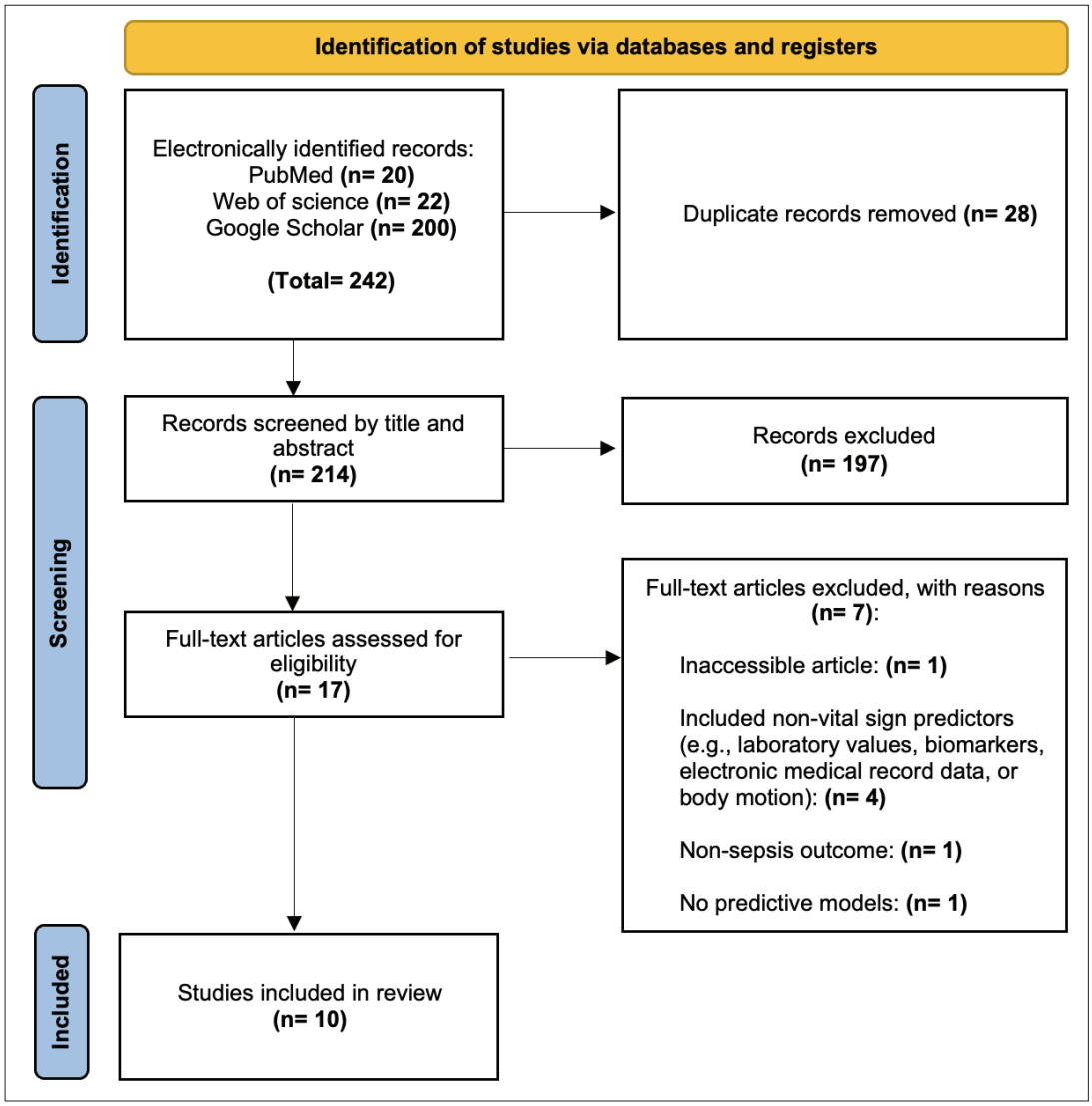



***Figure 1.*** *Preferred reporting items for systematic reviews and meta-analysis (PRISMA) diagram.*

modeling in each study. Additionally, we detailed the sepsis definitions in Table S1. Disagreements were resolved through discussion.

### *Quality assessment*

We utilized PROBAST tool, a specialized tool to assess studies that applied predictive modeling, consisting mainly of four domains to assess the quality in the included studies. It categorizes the risk of bias into three classes: high, low, and unclear risk of bias. First domain assesses if there is a risk of bias introduced by selection of participants, or if there are concerns that the included participants and setting do not match the review question. Second domain assesses if there is a risk of bias introduced by predictors or their assessment, or if there are concerns that the definition, assessment or timing of predictors in the model do not match the review question. The third domain assesses if there is a risk of bias introduced by the outcome or its determination, or if there are concerns that the outcome, its definition, timing or determination do not match the review question. The last domain assesses if there is a risk of bias introduced by the analysis. Any study with only one domain of high-risk bias will be considered to carry a high risk of bias [20].

### *Data analysis*

The reported AUC-ROC per model in each study was pooled as raw means with their corresponding 95% confidence intervals (95% CIs). To evaluate statistical heterogeneity among studies, the chi-square test (also known as Cochran's Q Test) was employed. The chi-square statistic (Cochran Q) was then used to determine the I^2 Value. A P-value of chi-square <0.05 indicated a significant heterogeneity. High heterogeneity was defined as an I^2 value ≥50%. We used R 4.5.2, the meta, dmetar, metafor, metaforest, dplyr, ggplot2, and readxl packages for data analysis. We used both fixed- and random-effect models based on the heterogeneity of the results. A P-value for the Z-score <0.05 was considered statistically significant. We conducted a subgroup analysis based on the deployed models. Due to the multilevel nature of our data, we performed a three-level meta-analysis to assess the variance among different levels. Information Criteria

(AIC and BIC) were used to assess the fitness of the built model. A quantitative synthesis was conducted to summarize the main findings of the models included in the present studies.

## Results

### *Literature search*

#### *PRISMA flow chart*

A total of 242 records were identified through electronic database searches (PubMed, Web of Science, and Google Scholar). After removal of duplicates, 214 records remained and were screened by title and abstract. Of these, 197 were excluded due to irrelevance to the scope of the research. The full texts of 17 articles were assessed for eligibility. Seven articles were excluded. Four of them were excluded due to the inclusion of non-vital signs predictors, and the remaining three were excluded due to inaccessibility, lack of predictive models, and involvement of non-sepsis outcome. Ultimately, ten studies were included in both the meta-analysis and the qualitative data synthesis (Figure 1).

### *Characteristics of the included studies*

A total of 10 studies encompassed a total of 8882 patients. Only two studies were prospective cohort studies, while others were retrospective cohort studies. Three studies were conducted in the Netherlands, while the remaining were conducted in Spain, Australia, France, Iran, India, Korea, and USA. The mean birth weight ranged from 860 to 3360 grams (Table 1). Half of the included studies were of high risk of bias due to unclear definitions of the outcome of interest, and analysis bias due to feature selection methods, while the remaining half of the studies were of low risk of bias (Table S2).

### *Meta analysis*

#### *Best performance models*

We pooled the AUC-ROC of the best performing models across studies. The best performing model was selected

***Table 1.*** *Characteristics of the included studies.*

| Study ID | Study characteristics | | | | Patient demographics | | | | | | What does the model measure? | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | aCountry | Study type | sample size | | Birth weight (grams) | | Gestational age (weeks), mean ± SD | | Male, *N* (%) | | Onset of sepsis (e.g. 6-12 hours) | Time window |
| | | | LOS | control | Patients | controls | patients | controls | patients | controls | | |
| *Kausch [21]* | *US* | *Retrospective study* | *2494* | | *1011 ± 298* | | *27.8 ± 2.9* | | *1,248 (50)* | | *24 hours Before the late-onset sepsis diagnosis* | *Within 24 hours* |
| *Nidamanuri [22]* | *India* | *Retrospective study* | *199* | *391* | *range (680-3600)* | | *range (24-41)* | | *-* | | *6–12 (hours before diagnosis)* | *6-12 hours* |
| *Yang [23]* | *Netherland* | *Retrospective study* | *51* | *68* | *1048.3 ± 307.4* | *1250.7 ± 247.4* | *27.93 ± 1.85* | *29.62 ± 1.48* | *29 (56.9)* | *29 (42.6)* | *6 hours before CRASH* | *-* |
| *Garstman [24]* | *Netherland* | *Retrospective study* | *14* | *25* | *2793 ± 185* | *1203.3 ± 688.9* | *28.33 ± 5.19* | *32.67 ± 6.67* | *3(21.4)* | *21(84)* | *48 hours before diagnosis* | *–48 hours* |
| *Hu [25]* | *Australia* | *Retrospective study* | | *146* | *-* | | *NR, BUT all are PRETERM in NICU* | *-* | *-* | *-* | *The model predicts the risk of late-onset neonatal sepsis within 24 hours based on vital sign patterns (24h before clinical suspicion)* | *t0-24 hours (t0 time point when antibiotic is ordered)* |
| *Gomez [26]* | *Spain* | *Prospective cohort study* | *15* | *64* | *3419 ± 373* | *3360 ± 536* | *39 ± 1* | *39 ± 1* | *10(66)* | *32(50)* | *The outcome was: sepsis vs no sepsis (sepsis risk within the first 48 hours of life)* | *-* |
| *Song [27]* | *Korea* | *Retrospective study* | *21* | *2798* | *860 ± 270* | *2040 ± 700* | *30.5 ± 5.56* | *34.0 ± 0.74* | *13 (61.9)* | *1508(53.9)* | *A prediction model that predicts LONS occurrence 48 hours earlier* | *48 hours* |

| Study ID | Study characteristics | | | | Patient demographics | | | | | | What does the model measure? | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | aCountry | Study type | sample size | | Birth weight (grams) | | Gestational age (weeks), mean ± SD | | Male, *N* (%) | | Onset of sepsis (e.g. 6-12 hours) | Time window |
| | | | LOS | control | Patients | controls | patients | controls | patients | controls | | |
| *León [28]* | *France* | *Prospective cohort study* | *24* | *25* | *868.3 ± 211.1* | | *26.6 ± 2.00* | | *15(62.5)* | *12(48)* | *KNN (calib 48 hours, ti30h) with graph features* | *6 hours before antibiotic start t = 0* |
| *Amiri [29]* | *Iran* | *Retrospective cohort study* | *14* | *14* | *-* | | *-* | | *-* | | *Overall observation period = 24 hours before clinical diagnosis (with key discriminative features identified ≈22 hours before diagnosis)* | *22 hours before clinical dx* |
| *Berg [30]* | *Netherland* | *Retrospective cohort study* | *389* | *2130* | *median (1020)* | *median (1300)* | *28.1 median* | *30.3 median* | *198/389 (50.9)* | *1147(53.8)* | *up to 24 hours before clinical suspension* | *–24 hours to +12 S32 hours around blood culture, AUC calculated at the time of blood culture, t = 0* |

from studies reporting multiple machine-learning models, based on the model selected by the study authors or the highest reported sensitivity and/or specificity. Using a random effects model, the pooled AUC was 0.84, 95% CI: [0.77, 0.92]. The results were extremely heterogeneous ($I^2$ = 99.2%, 12 = 0.007, Q = 1144.51, $p$ = 1.7e-246) (Figure 2).

### *Subgroup analysis based on the deployed model*

We performed a subgroup analysis based on the algorithm used in all studies. Random forest model (RF) and support vector machine (SVM) were the best performing models (AUC = 0.86,95% CI: [0.61; 1.11]), and (AUC: 0.86, 95% CI [0.64; 1.08]), respectively. K-nearest neighbors yielded a pooled AUC of 0.85, 95% CI: [0.61; 1.08]). Logistic regression (LR) and XGBoost yielded a comparable predictive ability (LR: 0.81, 95% CI: [0.76; 0.87]), and (AUC: 0.81, 95% CI: [0.70; 0.93]). Neural network models yielded a low predictive ability (AUC: 0.78, 95% CI: [0.75; 0.81]), and it was the only homogeneous subgroup ($p$ = 0.63, $I^2$ = 0%) (Figure 3).

## *Multilevel-meta-analysis*

The full three-level model yielded an AIC = –33.88, and BIC = –30.47, while between-studies variance yielded an AIC = –34.80, and BIC = –32.53. The heterogeneity in multilevel meta-analysis is driven almost entirely by the differences in effect sizes within the studies, rather than differences between the studies ($p$ = 2.99e-1). These findings indicate that the full three-level model is not necessary; a two-level model that accounts for within-study variance provides the best fit (Figure 4).

## *Qualitative synthesis of artificial intelligence architectures for early neonatal sepsis detection*

### *Overview of model development strategies*

The included studies used diverse ML and DL approaches. Performance depended on architecture, feature engineering, inputs, and prediction horizon. Models used continuously monitored vital signs and derived metrics. Multi-signal models outperformed single-signal ones, and strong feature engineering often yielded comparable performance to complex DL architectures.

## *Machine learning architectures*

### *Tree-based ensemble models*

Tree-based ensembles (RF, XGBoost, Gradient Boosting, AdaBoost) were the most consistently high-performing models, with AUROCs of 0.834 (XGBoost; Kausch et al.) to 0.973 (RF; Garstman et al.). AdaBoost reached 0.943 (Gomez et al.), Gradient Boosting 0.881 at 48 hours before diagnosis (Song et al.), and XGBoost 0.875 at 6 hours before onset (Yang et al.).

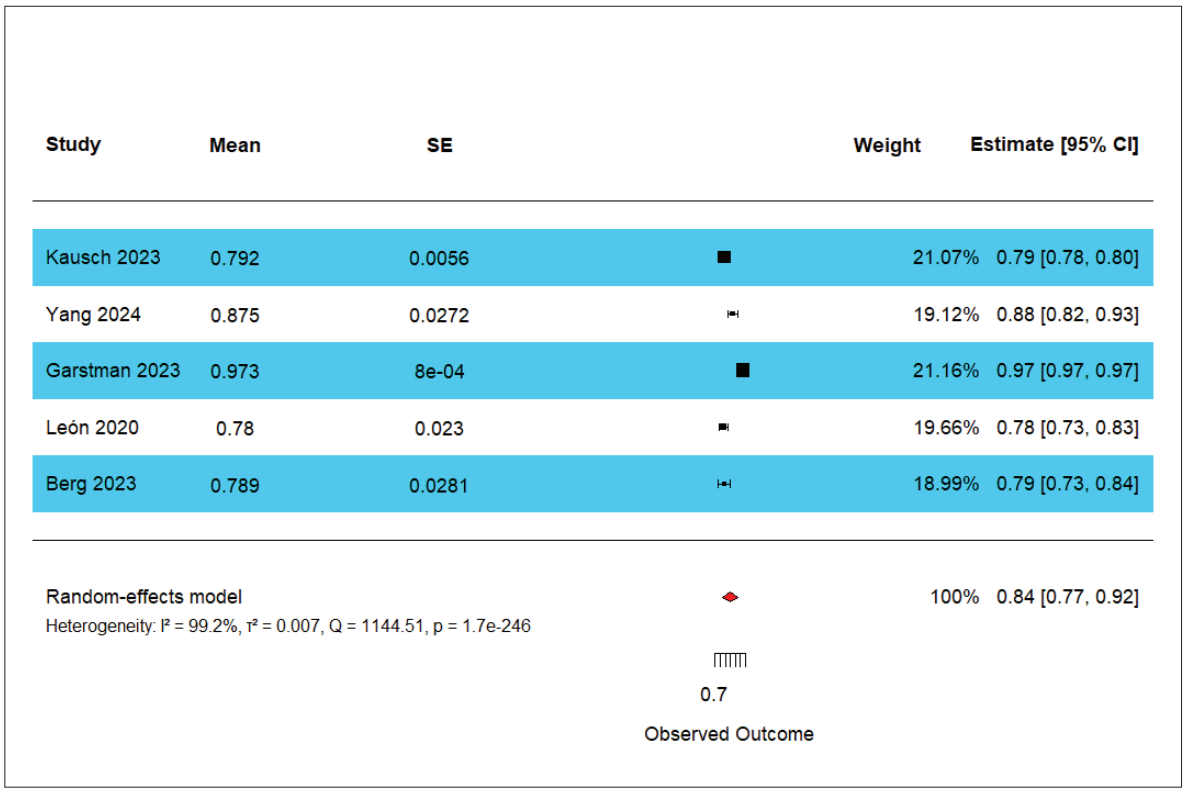



***Figure 2.*** *Best performance model of the included studies.*

### *Logistic regression and conventional statistical models*

Simpler statistical models remained competitive with well-engineered features. Logistic regression outperformed KNN, RF, and SVM when visibility graph features were added to HRV measures (AUROC 87.7%; León et al.), and the logistic regression-based POWS maintained stable external validation (AUROC 0.791-0.793; Kausch et al.). Logistic regression also reached AUROC 0.73 at clinical suspicion (Berg et al.). Their interpretability adds clinical value.

## ***Distance-based and kernel-based models***

SVMs and KNNs were evaluated in multiple studies and generally demonstrated strong but inconsistent performance.

Garstman et al. reported a KNN AUROC of 0.950, making it one of the highest-performing models while also exhibiting favorable computational efficiency. León et al. included both KNN and SVM models, although neither surpassed Logistic Regression when visibility graph features were incorporated. Yang et al. also evaluated KNN and SVM approaches but found XGBoost superior.

## ***Deep learning architectures***

### *Convolutional neural networks*

CNNs automated feature extraction from physiological time-series data. A 1D-CNN using 20 HRV features reached 88.9% accuracy about 22 hours before diagnosis (Amiri et al.). A fine-tuned GoogleNet applied to HR, RR, and $SpO^2$ signals converted into RGB images achieved an AUROC of 0.7903, lower than the leading ensemble models (Hu et al.). CNNs can identify temporal patterns of neonatal sepsis but do not consistently outperform optimized ML approaches.

### *Recurrent and hybrid deep learning models*

Hybrid architectures gave the strongest DL results: a CNN-LSTM outperformed RF, Gradient Boosting, and standalone LSTM (AUROC 0.93), identifying sepsis 6-12 hours before diagnosis (Nidamanuri et al.). Attention-based recurrent models did not outperform XGBoost (Yang et al.). Evidence that hybrid models surpass ensemble ML remains limited.

## ***Influence of feature engineering on model performance***

Feature engineering was one of the strongest determinants of predictive success. Visibility graph metrics improved AUROC by 6.8% (León et al.), 102 engineered HR, RR, and $SpO_2$ features achieved an AUROC of 0.875 (Yang et al.), and skewness and kurtosis enabled prediction up to 48 hours before diagnosis (Song et al.). Predictive value therefore depends not only on the algorithm but also on transforming raw signals into clinically meaningful representations.

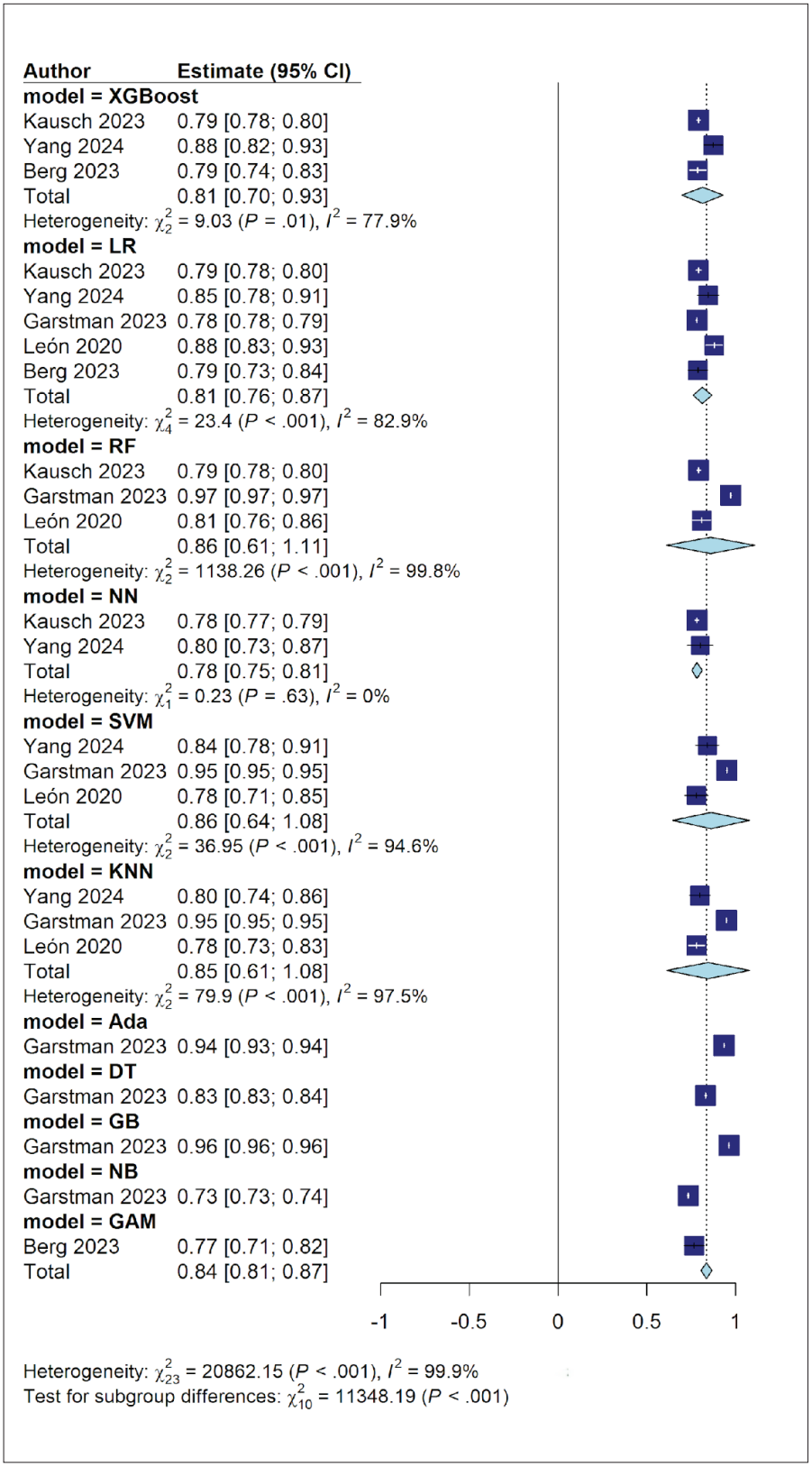



***Figure 3.*** *Subgroup analysis based on the deployed model in the included studies.*

```
> aov_table
                                                AIC       BIC        ll       LRT           p
Three-level model                          -33.88041 -30.47393 19.940207        NA          NA
Within-studies variance constrained        -24.05020 -21.77921 14.025100 11.830214 5.827729e-04
Between-studies variance constrained       -34.80571 -32.53472 19.402854  1.074706 2.998849e-01
Both variance components constrained       -17.65046 -16.51496  9.825229 20.229956 4.046885e-05
```

***Figure 4.*** *Multilevel-meta-analysis of the included studies.*

## Comparative assessment of architectures for early detection

Ensemble ML methods (RF, XGBoost, AdaBoost, Gradient Boosting) gave the most consistent results for both performance and early detection, often with AUROC above 0.85 and sometimes above 0.95. DL models, particularly CNN-LSTM, performed comparably but were evaluated in fewer studies and needed more extensive data streams. The earliest prediction horizons (22-48 hours before diagnosis) were achieved by advanced feature engineering combined with ensemble learning rather than DL alone, suggesting that better feature representation may contribute more than greater architectural complexity (Table 2).

## Discussion

This systematic review and meta-analysis shows that AI and ML models using continuous physiologic monitoring achieve clinically meaningful performance for LOS (pooled AUC 0.84, 95% CI: 0.77-0.92; 10 studies, 8,882 patients). RF and SVM performed best (AUC 0.86 each), whereas neural networks had the lowest AUC (0.78) but were the only homogeneous subgroup ($I^2 = 0\%$). To our knowledge, this is the first meta-analysis restricted to models trained on continuous vital signs, without mandatory laboratory, biomarker, or EHR inputs, providing an evidence base for non-invasive NICU surveillance.

Accurate prediction is essential for improving outcomes [31]. Blood culture, the gold standard, is positive in only about 50% of adults with clinical signs of sepsis, and less often after antimicrobials, yet adults still show more specific symptoms [9,32-36]. In neonates, sensitivity is lower (about 9% positive), largely because of small blood volumes, and time to positivity is 9-18 hours, sometimes up to 48 hours [37-41]. By contrast, some models in our review reached a sensitivity of up to 96% and could possibly predict sepsis 6 to 48 hours before clinical suspicion [21-30].

AI sepsis algorithms aim to improve accuracy and timeliness [42]. Adult models rely mainly on comprehensive EHR data, whereas NICU models can use the continuous signals of bedside monitors [43,44]. Our pooled AUC of 0.84 is comparable to the median reported for adult sepsis AI and requires no laboratory inputs, an advantage in resource-limited NICUs [45]. Such models may help bridge the diagnostic gap created by subtle, nonspecific neonatal symptoms [46].

The strong performance of RF is consistent with prior reviews, which found it the most frequently used algorithm and among the best performing [47]. Three deep learning studies excluded for dataset overlap support the potential

***Table 2.*** *Summary of findings of the included studies.*

| Study | Models used | Variables used | Key results |
|---|---|---|---|
| *Kausch [21]* | *Logistic Regression [Pulse Oximetry Warning System (POWS) model], Neural Network, XGBoost, and Random Forest.* | *HR and SpO2 statistics (mean, SD, skewness, kurtosis, cross-correlation).* | *XGBoost achieved the highest training AUC (0.834); POWS (LR) maintained strong external validation (AUC: 0.791–0.793).* |
| *Nidamanuri [22]* | *Random Forest, Gradient Boosting, Standalone LSTM, and hybrid CNN-LSTM.* | *Continuous HR, RR, SpO2, temperature, and blood pressure, plus derived HRV and apnea density metrics.* | *The CNN-LSTM model was most effective, achieving an AUROC of 0.93 and identifying sepsis 6–12 hours early.* |
| *Yang [23]* | *LR, KNN, SVM, XGB, Att-LSTM, and Att-ResNet.* | *HR, RR, and SpO2 sampled at 1 Hz, generating 102 features including visibility graph attributes.* | *Gradient Boosting (XGB) performed best with an AUC of 0.875 six hours before onset.* |
| *Garstman [24]* | *LogR, Naive Bayes (NB), Decision Trees (DT), Random Forest (RandF), KNN, SVM, ADA, and GB.* | *Features derived from HRV, Respiratory Frequency (RF), Perfusion Index (PFI), and Oxygen Saturation (SpO2).* | *Random Forest performed best (AUROC: 0.973); KNN was most efficient when considering training time (AUROC: 0.950).* |
| *Hu [25]* | *Convolutional Neural Networks (CNN) and fine-tuned GoogleNet.* | *Continuous HR, RR, and SpO2 data transformed into two-dimensional RGB images.* | *The fine-tuned GoogleNet model achieved an AUC of 0.7903.* |
| *Gomez [26]* | *AdaBoost, Bagged Classification Trees (BCT), RandF, LR, SVM, NB, CT, and k-NN.* | *HRV components (time, frequency, and non-linear).* | *AdaBoost was the superior model with an AUC of 0.943, followed by BCT (0.880) and Random Forest (0.840).* |
| *Song [27]* | *LR, Gaussian NB, DT, GB, AdaBoost, Bagging, RandF, and Multilayer Perceptron.* | *Non-invasive vital signs (HR, BP, SpO2, RR, temperature) processed for statistical features like kurtosis and skewness.* | *Gradient Boosting achieved an AUROC of 0.881 for prediction (48h early) and 0.877 for onset detection.* |
| *León [28]* | *KNN, Logistic Regression, Random Forest, and SVM.* | *Traditional HRV features combined with novel Visibility Graph (VG) indexes.* | *Logistic Regression using VG features performed best, reaching an AUROC of 87.7% in the six hours before treatment.* |
| *Amiri [29]* | *1D Convolutional Neural Network (1D-CNN).* | *20 selected Heart Rate Variability (HRV) features (time, frequency, and time-frequency domains), including VLF and LF.* | *Achieved 88.89 ± 7.86% overall accuracy for classification 22 hours prior to clinical diagnosis.* |
| *Berg [30]* | *Logistic Regression (LR), Generalized Additive Models (GAMs), and XGBoost.* | *Routinely measured Heart Rate (HR) and Oxygen Saturation (SpO2), including bradycardia/ tachycardia counts and saturation drops.* | *The LR model yielded an AUC of 0.73 at clinical suspicion; simulation detected sepsis in at least 47% of patients before suspicion.* |

of deep learning: Kallonen et al. reported an AUROC of 0.81 with a median 44-hour lead time, León et al. 90.4% over the six hours before antibiotics, and Peng et al. 0.77, improving closer to onset [48-50].

## Limitations

Several limitations apply. Most studies were retrospective, with only two prospective cohorts, which limits confidence in real-world applicability. Half were rated high risk of bias by PROBAST, mainly because of unclear sepsis outcome definitions and questionable feature selection, which may have inflated reported AUCs. The lack of a universal definition of LOS created heterogeneity in training targets and complicates cross-study comparison. Small sample sizes raise overfitting concerns. Finally, studies were concentrated in high-income countries, limiting generalizability to low- and middle-income settings, where the sepsis burden is highest.

## Clinical Implications

The ability of these models to detect LOS hours before clinical suspicion, using signals already available from standard NICU bedside monitors, provides a strong rationale for piloting them as decision support tools. These models provide a long interval between the time of prediction and definitive diagnosis, which offer an opportunity for early intervention that may reduce morbidity, unnecessary antibiotic exposure, and antimicrobial resistance.

## Future Research

Future research should prioritize large-scale, multicenter, prospective studies with harmonized sepsis definitions. Additionally, similar studies of AI and ML testing should be integrated into low- and middle-income countries, where laboratory diagnostics are limited, and the burden of neonatal sepsis is higher.

## Conclusion

AI and ML models using continuous physiologic monitoring may help predict LOS (pooled AUC 0.84) 6–48 hours before clinical diagnosis, with Random Forest and SVM performing best. However, substantial heterogeneity ($I^2 = 99.2\%$) warrants cautious interpretation and prospective validation before clinical use.


### Acknowledgements

None.


### List of Abbreviations

| | |
|---|---|
| 1D-CNN | One-dimensional convolutional neural network |
| ADA | Adaptive boosting |
| AI | Artificial intelligence |
| Att-LSTM | Attention-based long short-term memory neural network |
| Att-ResNet | Attention-based Residual convolutional neural network |
| AUC-ROC | Area-under the curve from receiver operating characteristics curve |
| BCT | Bagged Classification Trees. |
| BP | Blood pressure |
| CNN | Convolutional neural networks |
| CNN-LSTM | Convolutional neural network- Long short-term memory |
| DL | Deep Learning |
| DT | Decision trees |
| EHR | Electronic Health Record |
| GAMs | Generalized Additive Models |
| GBM | Gradient Boosting Machine |
| HR | Heart rate |
| HRV | Heart rate variability |
| KNN | K-nearest neighbors |
| LMICs | Low and middle-income countries |
| LOS | Late-onset neonatal sepsis |
| LR | Logistic regression |
| LSTM | Long short-term memory |
| ML | Machine learning |
| NB | Naive Bayes |
| NICU | Neonatal intensive care unit |
| NNM | Neural network models |
| NR | NOT Reported |
| PFI | Perfusion Index |
| POWS | Pulse Oximetry Warning System |
| RF | Random forest model. |
| RF | Respiratory Frequency |
| RNN | Recurrent neural network. |
| SpO2 | Oxygen Saturation |
| SVM | Support vector machine. |
| VG | Visibility Graph |
| VLBW | Very low birth weight |
| XGB | Extreme gradient boosting XGBoost |

### Conflict of interests

All authors declare that they have no conflict of interests.


### Funding

None.


### Ethical approval

Not applicable (systematic review, no human or animal data used).

### Data availability

All data are included in this article.


### Author details

Layla Abdullah Alajlan[1], Aseel M Alsamaani[1], Rola H Alseghair[1], Sarah A Alomar[1] , Hind H Almazyad[1], Razan A Aljameeli[1], Abdulrahman Obaid Alharbi[2]

1. College of Medicine, Qassim University, Buraydah, Qassim, Saudi Arabia
2. Neonatologist and Head of the Neonatal Intensive Care Service Line, Qassim Health Cluster, Buraydah, Qassim, Saudi Arabia


*Supplementary content (If any) is available online.*

### References


1. Grassham CD, Kamm C, Pao aroon CN, McGuire KN, Maxwell JR. Improving outcomes in neonatal late onset sepsis: where every second matters. J Neonatal

Nurs. 2020;27(2):100–5. https://doi.org/10.1016/j.jnn.2020.07.011

2. Kaffe K, Syrogiannopoulos GA, Petinaki E, Goudesidou M, Kalaitzi A, Gounaris A, et al. Epidemiology and outcomes of late-onset neonatal sepsis in preterm infants in a tertiary hospital. Children (Basel). 2025;12(5):532. https://doi.org/10.3390/children12050532

3. Afonso E, Smets K, Deschepper M, Verstraete E, Blot S. The effect of late-onset sepsis on mortality across different gestational ages in a neonatal intensive care unit: a historical study. Intensive Crit Care Nurs. 2023;77:103421. https://doi.org/10.1016/j.iccn.2023.103421

4. Russell N, Barday M, Okomo U, Dramowski A, Sharland M, Bekker A. Early-versus late-onset sepsis in neonates-time to shift the paradigm? Clin Microbiol Infect. 2024;30(1):38–43. https://doi.org/10.1016/j.cmi.2023.07.023

5. Herbozo C, Julca I, Flores F, Hernandez R, Zegarra J. Incidence and microbiological characteristics of neonatal late onset sepsis in a neonatal intensive care unit in peru. Int J Infect Dis. 2021;108:171–5. https://doi.org/10.1016/j.ijid.2021.05.012

6. Guan H, Liang X, Liang Y, Yang F. Global trends in neonatal sepsis mortality and attributable risk factors from 1990 to 2021. J Paediatr Child Health. 2026;62(1):106–17. https://doi.org/10.1111/jpc.70241

7. Mackay CA, Nathan EA, Porter MC, Shrestha D, Kohan R, Strunk T. Epidemiology and outcomes of neonatal sepsis: experience from a tertiary Australian NICU. Neonatology. 2024;121(6):703–14. https://doi.org/10.1159/000539174

8. Ng S, Strunk T, Jiang P, Muk T, Sangild PT, Currie A. Precision medicine for neonatal sepsis. Front Mol Biosci. 2018;5:70. https://doi.org/10.3389/fmolb.2018.00070

9. Zea-Vera A, Ochoa TJ. Challenges in the diagnosis and management of neonatal sepsis. J Trop Pediatr. 2015;61(1):1–13. https://doi.org/10.1093/tropej/fmu079

10. Kariniotaki C, Thomou C, Gkentzi D, Panteris E, Dimitriou G, Hatzidaki E. Neonatal sepsis: a comprehensive review. Antibiotics (Basel). 2024;14(1):6. https://doi.org/10.3390/antibiotics14010006

11. Wynn JL. Defining neonatal sepsis. Curr Opin Pediatr. 2016;28(2):135–40. https://doi.org/10.1097/mop.0000000000000315

12. Cantey JB, Patel SJ. Antimicrobial stewardship in the NICU. Infect Dis Clin North Am. 2014;28(2):247–61. https://doi.org/10.1016/j.idc.2014.01.005

13. Ranjeva SL, Warf BC, Schiff SJ. Economic burden of neonatal sepsis in sub-saharan africa. BMJ Glob Health. 2018;3(1):e000347. https://doi.org/10.1136/bmjgh-2017-000347

14. Salman O, Procter SR, McGregor C, Paul P, Hutubessy R, Lawn JE, et al. Systematic review on the acute cost-of-illness of sepsis and meningitis in neonates and infants. Pediatr Infect Dis J. 2020;39(1):35–40. https://doi.org/10.1097/inf.0000000000002500

15. Sullivan BA, Kausch SL, Fairchild KD. Artificial and human intelligence for early identification of neonatal sepsis. Pediatr Res. 2023;93(2):350–6. https://doi.org/10.1038/s41390-022-02274-7

16. Fairchild KD. Predictive monitoring for early detection of sepsis in neonatal ICU patients. Curr Opin Pediatr. 2013;25(2):172–9. https://doi.org/10.1097/mop.0b013e32835e8fe6

17. Neal SR, Sturrock SS, Musorowegomo D, Gannon H, Zaman M, Cortina-Borja M, et al. Clinical prediction models to diagnose neonatal sepsis in low-income and middle-income countries: a scoping review. BMJ Glob Health. 2025;10(4):e017582. https://doi.org/10.1136/bmjgh-2024-017582

18. Tudor S, Bhatia R, Liem M, Wani TA, Boyd J, Raza Khan U. Opportunities and challenges of using artificial intelligence in predicting clinical outcomes and length of stay in neonatal intensive care units: systematic review. J Med Internet Res. 2025;27:e63175. https://doi.org/10.2196/63175

19. Tetzlaff J, Page M, Moher D. Pns154 the prisma 2020 statement: development of and key changes in an updated guideline for reporting systematic reviews and meta-analyses. Value Health. 2020;23:S312–3. https://doi.org/10.1016/j.jval.2020.04.1154

20. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. https://doi.org/10.7326/m18-1376

21. Kausch SL, Brandberg JG, Qiu J, Panda A, Binai A, Isler J, et al. Cardiorespiratory signature of neonatal sepsis: development and validation of prediction models in 3 nicus. Pediatr Res. 2023;93(7):1913–21. https://doi.org/10.1038/s41390-022-02444-7

22. Nidamanuri S, Bhosale YH, Jijaba KJ, Kumar BK, Malathi L. Neonatal sepsis prediction using AI-based vital sign analysis in nicus: a multicenter retrospective study. Vasc Endovasc Rev. 2025;8(17s):375–82.

23. Yang M, Peng Z, van Pul C, Andriessen P, Dong K, Silvertand D, et al. Continuous prediction and clinical alarm management of late-onset sepsis in preterm infants using vital signs from a patient monitor. Comput Methods Programs Biomed. 2024;255:108335. https://doi.org/10.1016/j.cmpb.2024.108335

24. Garstman AG, Rodriguez Rivero C, Onland W. Early detection of late onset sepsis in extremely preterm infants using machine learning: towards an early warning system. Appl Sci. 2023;13(16):9049. https://doi.org/10.3390/app13169049

25. Hu Y, Lee VC, Tan K. An application of convolutional neural networks for the early detection of late-onset neonatal sepsis. In: 2019 International Joint Conference on neural networks (IJCNN). IEEE; 2019. pp. 1–8. https://doi.org/10.1109/ijcnn.2019.8851683

26. Gomez R, Garcia N, Collantes G, Ponce F, Redon P. Development of a non-invasive procedure to early detect neonatal sepsis using HRV monitoring and machine learning algorithms. In: IEEE international symposium on computer-based medical systems (CBMS). IEEE; 2019. pp. 132–7. https://doi.org/10.1109/cbms.2019.00037

27. Song W, Jung SY, Baek H, Choi CW, Jung YH, Yoo S. A predictive model based on machine learning for the early detection of late-onset neonatal sepsis: development and observational study. JMIR Med Inform. 2020;8(7):e15965. https://doi.org/10.2196/15965

28. Leon C, Carrault G, Pladys P, Beuchee A. Early detection of late onset sepsis in premature infants using visibility graph analysis of heart rate variability. IEEE J Biomed Health

Inform. 2021;25(4):1006–17. https://doi.org/10.1109/jbhi.2020.3021662

29. Amiri P, Abbasi H, Derakhshan A, Gharib B, Nooralishahi B, Mirzaaghayan M. Potential prognostic markers in the heart rate variability features for early diagnosis of sepsis in the pediatric intensive care unit using convolutional neural network classifiers. In: Annual international conference of the IEEE engineering in medicine and biology society. IEEE engineering in medicine and biology society. Annual international conference. IEEE; 2020 July. pp. 5627–5630. https://doi.org/10.1109/embc44109.2020.9175481

30. van den Berg M, Medina O, Loohuis I, van der Flier M, Dudink J, Benders M, et al. Development and clinical impact assessment of a machine-learning model for early prediction of late-onset sepsis. Comput Biol Med. 2023;163:107156. https://doi.org/10.1016/j.compbiomed.2023.107156

31. Møller MH, Alhazzani W, Shankar-Hari M. Focus on sepsis. Intensive Care Med. 2019;45(10):1459-1461. https://doi.org/10.1007/s00134-019-05680-4

32. Brun-Buisson C. Incidence, risk factors, and outcome of severe sepsis and septic shock in adults: a multicenter prospective study in intensive care units. JAMA. 1995;274(12):968. https://doi.org/10.1001/jama.1995.03530120060042

33. Grace CJ, Lieberman J, Pierce K, Littenberg B. Usefulness of blood culture for hospitalized patients who are receiving antibiotic therapy. Clin Infect Dis. 2001;32(11):1651–5. https://doi.org/10.1086/320527

34. Scheer CS, Fuchs C, Gründling M, Vollmer M, Bast J, Bohnert JA, et al. Impact of antibiotic administration on blood culture positivity at the beginning of sepsis: a prospective clinical cohort study. Clin Microbiol Infect. 2019;25(3):326–31. https://doi.org/10.1016/j.cmi.2018.05.016

35. Cheng MP, Stenstrom R, Paquette K, Stabler SN, Akhter M, Davidson AC, et al. Blood culture results before and after antimicrobial administration in patients with severe manifestations of sepsis: a diagnostic study. Ann Intern Med. 2019;171(8):547–54. https://doi.org/10.7326/m19-1696

36. Pazin GJ, Saul S, Thompson ME. Blood culture positivity: suppression by outpatient antibiotic therapy in patients with bacterial endocarditis. Archives of Internal Medicine. 1982;142(2):263–8. https://doi.org/10.1001/archinte.1982.00340150063012

37. Schelonka RL, Chai MK, Yoder BA, Hensley D, Brockett RM, Ascher DP. Volume of blood required to detect common neonatal pathogens. J Pediatr. 1996;129(2):275–8. https://doi.org/10.1016/s0022-3476(96)70254-8

38. Kosmeri C, Giapros V, Serbis A, Baltogianni M. Application of advanced molecular methods to study early-onset neonatal sepsis. Int J Mol Sci. 2024;25(4):2258. https://doi.org/10.3390/ijms25042258

39. Shashinika KL, Mendis IS, Gayan T, Rathnayake AM, Waidyasuriya LS, Ambillapitiya RL, et al. Evaluation of clinical and laboratory parameters for recognition and diagnosis of neonatal sepsis: a retrospective analysis of a cross-sectional study. Pediatr Med. 2025;8:24. https://doi.org/10.21037/pm-25-14

40. Wynn JL, Polin RA. Progress in the management of neonatal sepsis: the importance of a consensus definition. Pediatr Res. 2018;83(1-1):13–5. https://doi.org/10.1038/pr.2017.224

41. Klingenberg C, Kornelisse RF, Buonocore G, Maier RF, Stocker M. Culture-negative early-onset neonatal sepsis-at the crossroad between efficient sepsis care and antimicrobial stewardship. Front Pediatr. 2018;6:285. https://doi.org/10.3389/fped.2018.00285

42. Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. Prediction of sepsis patients using machine learning approach: a meta-analysis. Comput Methods Programs Biomed. 2019;170:1–9. https://doi.org/10.1016/j.cmpb.2018.12.027

43. Haas R, McGill SC. Artificial intelligence for the prediction of sepsis in adults. Can J Health Technol. 2022;2(3). https://doi.org/10.51731/cjht.2022.283

44. Kumar N, Akangire G, Sullivan B, Fairchild K, Sampath V. Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront. Pediatr Res. 2020;87(2):210–220. https://doi.org/10.1038/s41390-019-0527-0

45. Abbas GH, Sen P, Giri OA, Khan NH. Artificial intelligence-based predictive modeling for early detection of sepsis in hospitalized patients: a systematic review and meta-analysis. Crit Care Explor. 2025;7(12):e1360. https://doi.org/10.1097/cce.0000000000001360

46. Li F, Wang S, Gao Z, Qing M, Pan S, Liu Y, et al. Harnessing artificial intelligence in sepsis care: advances in early detection, personalized treatment, and real-time monitoring. Front Med (Lausanne). 2024;11:1510792. https://doi.org/10.3389/fmed.2024.1510792

47. Kainth D, Prakash S, Sankar MJ. Diagnostic performance of machine learning-based models in neonatal sepsis: a systematic review. Pediatr Infect Dis J. 2024;43(9):889–901. https://doi.org/10.1097/inf.0000000000004409

48. Kallonen A, Juutinen M, Värri A, Carrault G, Pladys P, Beuchée A. Early detection of late-onset neonatal sepsis from noninvasive biosignals using deep learning: a multicenter prospective development and validation study. Int J Med Inform. 2024;184:105366. https://doi.org/10.1016/j.ijmedinf.2024.105366

49. León C, Pladys P, Beuchée A, Carrault G. Recurrent neural networks for early detection of late onset sepsis in premature infants using heart rate variability. In: Computing in cardiology. IEEE Computer Society; 2021. pp. 1–4. https://doi.org/10.23919/cinc53138.2021.9662715

50. Peng Z, Varisco G, Liang RH, Kommers D, Cottaar W, Andriessen P, et al. DeepLOS: deep learning for late-onset sepsis prediction in preterm infants using heart rate variability. Smart Health. 2022;26:100335. https://doi.org/10.1016/j.smhl.2022.100335