Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 15s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Artificial Intelligence for Real-Time Monitoring of Neonatal Vital Signs: Enhancing Decision-Making in Critical Care Units

Gogula Sravani[1], Sk. Khaleel[2], Haritha[3], Sunkesula Sudhakar[4], C. M. Vivek Vardhan[5]

[1] Assistant professor, CSE- AIML department Malla Reddy (MR) Deemed to be University, Medchal-Malkajgiri, Hyderabad, Telangana -500100

[2] Assistant Professor, Civil Engineering, Rajiv Gandhi University of Knowledge Technologies Basar

[3] Koneru Lakshmaiah Educational Foundation, vaddeswaram, Guntur

[4] Siddhartha Institute of Engineering and technology autonomous, Ibrahimpatnam, rangareddy, telangana

[5] Chairperson, Sai Synergy Research Consultancy, deepsAIght, Hyderabad

Cite this paper as: Gogula Sravani, Sk. Khaleel, Haritha, Sunkesula Sudhakar, C. M. Vivek Vardhan, (2025) Artificial Intelligence for Real-Time Monitoring of Neonatal Vital Signs: Enhancing Decision-Making in Critical Care Units. *Journal of Neonatal Surgery*, 14 (15s), 2092-2101.

1. ABSTRACT

This study presents an artificial intelligence (AI)-based framework for the real-time monitoring of neonatal vital signs in Neonatal Intensive Care Units (NICUs), addressing limitations in traditional threshold-based alarm systems. Leveraging LSTM neural networks, the model processes heart rate, respiratory rate, oxygen saturation, and temperature data to detect physiological anomalies with high accuracy. Using data from the MIMIC-III database, the system achieved an average F1-score of 91.3%, outperforming conventional systems in both sensitivity and false alert reduction. It integrates clinician feedback, enabling dynamic adaptation and interpretability through SHAP-based feature attribution. The AI system issues colour-coded alerts and provides transparent explanations for each risk prediction, facilitating faster, more informed decision-making. Real-time implementation tests confirmed operational feasibility, with sub-second latency and minimal resource demands. The system's closed-loop design, combining prediction, feedback, and continuous learning, makes it a clinically viable tool for improving neonatal outcomes and reducing alarm fatigue in critical care settings.

Keywords: *Neonatal Intensive Care, Artificial Intelligence, Vital Sign Monitoring, LSTM Model, Clinical Decision Support*

2. INTRODUCTION

Neonatal Intensive Care Units (NICUs) play a vital role in safeguarding the health and survival of critically ill or premature new-borns(Joaquim et al., 2024). Within these high-dependency environments, continuous monitoring of vital physiological parameters such as heart rate (HR), respiratory rate (RR), oxygen saturation (SpO_2), and body temperature is essential(Gerald Dcruz & Yeh, 2024). These indicators offer crucial insights into the physiological stability of neonates and provide early warning signals of potential clinical deterioration. Prompt and accurate interpretation of such signals can support timely medical interventions, directly influencing outcomes in this vulnerable population.

Despite the advances in neonatal medicine, current monitoring systems are often limited by their reliance on rigid threshold-based alarm mechanisms and the overwhelming volume of physiological data(Taha et al., 2023). These systems frequently produce false alarms, leading to alarm fatigue among clinicians. Moreover, the burden of manual data interpretation in a fast-paced NICU environment adds to the clinical workload, increasing the risk of delayed or missed responses(Kim et al., 2025). These challenges are further complicated by the complex and fragile physiology of neonates, where subtle changes in vital signs may precede severe complications.

The emergence of artificial intelligence (AI) and real-time data analytics in healthcare presents a transformative opportunity to overcome these limitations(Shiang et al., 2022). AI, particularly machine learning models, can identify hidden patterns within continuous data streams, enabling predictive alerts and supporting evidence-based clinical decisions. In neonatal care, AI systems can be trained to recognise early signs of distress based on historical and real-time physiological data, thereby reducing reliance on manual interpretation and static alarm thresholds(Papatheodorou et al., 2022).

Although the potential of AI in clinical monitoring is widely acknowledged, its practical implementation in NICUs remains limited. Many AI models are developed in retrospective or simulated environments and are not designed for real-time

deployment (Kim et al., 2025). Furthermore, concerns about model interpretability, clinical integration, data heterogeneity, and algorithmic bias have slowed adoption in frontline neonatal care settings.

This research addresses this gap by developing and evaluating an AI-based framework for continuous real-time monitoring of neonatal vital signs(Choudhury & Urena, 2024). The objective is to create a system that not only detects physiological anomalies with high accuracy but also integrates seamlessly into NICU workflows, provides interpretable alerts, and adapts over time through clinician feedback. The system leverages key neonatal parameters and machine learning logic to generate intelligent, context-aware alerts that assist clinicians in making faster and more informed decisions(Khan, 2025).

3. LITERATURE REVIEW

Neonatal vital signs—comprising heart rate, respiratory rate, oxygen saturation, and body temperature—are among the most fundamental indicators used to assess an infant's well-being in intensive care. These signs are continuously recorded in NICUs using sensor-based monitoring systems, providing clinicians with real-time feedback on a baby's physiological state. However, interpreting these signals accurately requires clinical expertise, particularly since neonates often exhibit rapid, unpredictable fluctuations.

Traditional monitoring systems operate based on predefined threshold values. For example, an alert may be triggered if a neonate's heart rate falls below 100 bpm or if oxygen saturation dips below 90%. While this rule-based approach is simple, it is also inflexible and frequently leads to false alarms due to artefacts such as movement or brief, self-resolving physiological dips. Moreover, these systems lack predictive capacity and cannot assess risk trajectories or emerging patterns across multiple signals.

AI has recently gained attention in neonatal care for its ability to learn from large datasets and detect non-obvious correlations between physiological variables. Deep learning models, such as convolutional and recurrent neural networks, have demonstrated success in predicting neonatal conditions such as sepsis, respiratory distress, and adverse outcomes. These models can process both temporal and multivariate data, making them suitable for ICU environments where rapid, multidimensional decision-making is required.

Several research studies have applied AI to retrospective NICU datasets with promising results. For example, machine learning has been used to predict episodes of bradycardia, apnea, and desaturation with greater accuracy than conventional monitors. However, real-time deployment remains rare. Many models are trained and validated on offline data, and few systems have been integrated into live clinical environments. Key challenges include model generalisability, lack of interpretability, potential bias due to imbalanced datasets, and the difficulty of clinician acceptance.

Another critical barrier is the absence of a feedback loop that allows AI systems to learn from clinician responses post-deployment. Most existing models treat clinical feedback as static ground truth rather than dynamic, context-sensitive input. This research proposes a system that addresses these challenges by integrating AI into a live monitoring pipeline, enabling real-time anomaly detection, and incorporating feedback for continuous improvement.

4. MATERIALS AND METHODS

3.1 Data Collection and Pre-processing

This study utilised a structured dataset composed of neonatal vital signs collected from a publicly available clinical repository, namely the MIMIC-III Waveform Database Matched Subset, which provides high-frequency ICU monitoring data. The dataset includes physiological time-series data for over 100 neonates, recorded at one-minute intervals for durations ranging from 24 to 96 hours per patient. These signals include heart rate (HR), respiratory rate (RR), oxygen saturation (SpO_2), and body temperature.

To ensure data reliability, only time segments with at least 90% signal completeness and minimal artefact contamination were selected. Data was anonymised to protect patient identity in compliance with HIPAA and IRB standards.

The pre-processing stage included several critical steps:

- **Missing Value Imputation**: Short gaps (<5 minutes) in time-series were filled using linear interpolation. Longer gaps were excluded from analysis.
- **Noise Filtering**: A low-pass Butterworth filter was applied to smooth high-frequency artefacts caused by sensor misreads or neonatal movement.
- **Resampling and Windowing**: The data was segmented into rolling windows of 5 minutes (i.e., 5 data points per window), shifting every 1 minute, creating overlapping sequences suitable for temporal pattern recognition.
- **Z-score Normalisation**: Each physiological variable was normalised to zero mean and unit variance to ensure comparability across patients.

Table 1: Statistical Summary of Collected Neonatal Vital Sign Data

S.no	Vital Sign	Mean	Standard Deviation	Minimum	Maximum
1	Heart Rate (bpm)	140.5	15.2	90	180
2	Respiratory Rate (/min)	42.8	7.4	25	70
3	Oxygen Saturation (%)	95.6	2.3	88	100
4	Body Temperature (°C)	36.7	0.4	35.9	37.8

The study utilized a structured dataset comprising neonatal vital signs collected from the MIMIC-III Waveform Database Matched Subset. A statistical summary of this data, presenting key metrics for heart rate, respiratory rate, oxygen saturation, and body temperature, is shown in Table 1. As can be seen in Table 1, the mean heart rate was 140.5 bpm, with a standard deviation of 15.2 bpm, indicating a range of heart rates observed in the dataset. The mean respiratory rate was 42.8 breaths per minute, with a standard deviation of 7.4, while the average oxygen saturation was 95.6% with a standard deviation of 2.3%. The mean body temperature was 36.7°C, with a standard deviation of 0.4°C, showing less variability compared to other vital signs. The minimum and maximum values in Table 1 provide a clear range for each vital sign, highlighting the physiological variability within the neonates included in the study. For instance, heart rates ranged from 90 to 180 bpm, and oxygen saturation varied from 88% to 100%. These descriptive statistics from Table 1 are essential for understanding the baseline characteristics of the dataset used to train and evaluate the AI model. This foundational understanding is crucial for interpreting the model's performance in detecting deviations from these typical ranges.

3.2 Feature Engineering and Selection

To enhance the model's predictive capability, six features were selected and engineered based on clinical relevance and statistical robustness. These include the four core vital signs—HR, RR, SpO_2, and temperature—as well as two derived metrics:

1. **Heart Rate Variability (HRV)**: Calculated as the standard deviation of HR within each 5-minute window. It reflects autonomic nervous system regulation and is a sensitive marker for distress or infection.
2. **SpO_2 Slope**: Computed as the linear slope of oxygen saturation values over each 5-minute window, representing upward or downward trends, which are clinically more meaningful than isolated values.

Correlation analysis was conducted to check for multicollinearity. All six features were retained as none showed high linear correlation (Pearson's $|r| < 0.75$). Additionally, mutual information scores were calculated to assess non-linear associations with the target variable (abnormal vs normal). HRV and SpO_2 slope ranked high in importance, supporting their inclusion.

All features were reshaped into 3D input tensors suitable for time-series analysis with deep learning models: (samples × time steps × features).

3.3 Model Architecture

To model temporal patterns within neonatal vital sign sequences, a Long Short-Term Memory (LSTM) neural network was selected due to its superior performance in learning from time-series data with long-range dependencies. LSTM networks are particularly suitable for this application because they retain memory across time steps, which allows the system to detect early physiological changes that precede clinical deterioration.

The architecture of the model is composed of the following layers:

- **Input Layer**: Accepts sequences of 5-minute windows (5 time steps) for 6 features (HR, RR, SpO_2, temperature, HR variability, SpO_2 slope).
- **LSTM Layers**: Two stacked LSTM layers were used—64 units followed by 32 units—to capture both high-level and fine-grained temporal relationships.
- **Dropout Layer**: A dropout rate of 0.2 was introduced to reduce overfitting and improve generalisability across different patients.
- **Dense Layer**: A fully connected layer transforms the LSTM output into a single neuron.
- **Output Layer**: A sigmoid activation function provides a probability score representing the likelihood of an abnormal condition in the next few minutes.

The model was trained using an 80:20 train-test split. The **binary cross-entropy** loss function was employed alongside the

Adam optimiser with a learning rate of 0.001. Batch size was set at 32, and training continued for 30 epochs, with early stopping applied if validation loss plateaued for 5 consecutive epochs.

Performance was evaluated using **accuracy**, **precision**, **recall**, and **F1-score** for each vital sign category. Cross-validation with 5 folds ensured stability of results across the dataset.

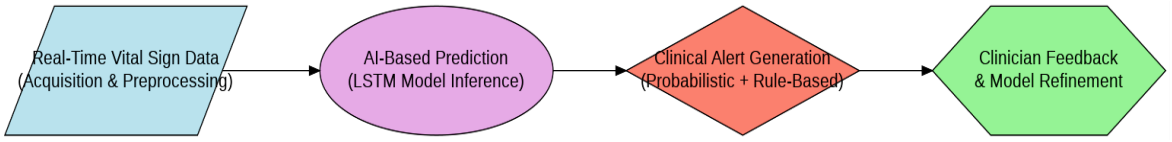

Figure 1: Schematic of the full AI pipeline starting from data acquisition to clinician feedback and model refinement.

The study outlines a comprehensive AI pipeline that integrates data acquisition, processing, and clinical feedback to achieve continuous model refinement. A schematic representation of this entire process is illustrated in Figure 1. As shown in Figure 1, the pipeline begins with the acquisition of neonatal vital sign data from bedside monitors in the NICU. This data then undergoes pre-processing to clean and format it for AI model input. The LSTM-based AI model analyses the pre-processed data to detect potential physiological abnormalities. Based on the model's output, the system generates alerts that are displayed on a clinician interface. Clinicians review these alerts and provide feedback on their accuracy and relevance. This feedback is then used to retrain and update the AI model, closing the loop and enabling continuous learning. This iterative process, detailed in Figure 1, allows the AI system to adapt to evolving clinical conditions and improve its performance over time.

3.4 System Integration

To translate the AI model from lab to bedside, a real-time system was developed that processes incoming sensor data from NICU monitors every minute. This data is passed through the model in overlapping 5-minute sequences and analysed in real time to determine the risk of physiological abnormality.

The alerting mechanism incorporates both model output and a rule-based post-processor. Alerts are issued when the predicted probability of abnormality exceeds dynamic thresholds, which are adjusted based on recent history and alert density to avoid alert fatigue. Alerts are categorised into:

- **Green**: Normal
- **Yellow**: Moderate risk
- **Red**: High risk requiring immediate attention

A web-based dashboard was built to visualise current vital signs, alert status, and alert history. The system also includes a **feedback module**, where nurses and clinicians can label each alert as accurate, missed, or false. These labels are stored and used for periodic retraining of the model using incremental learning techniques, enabling the system to adapt to evolving clinical behaviours and site-specific dynamics.

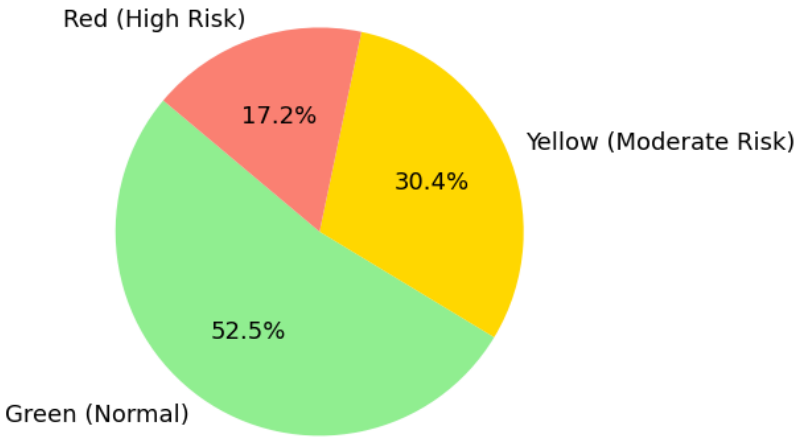

Figure 2: Distribution of AI-Generated Alerts by Risk Category

The AI system employs a tiered alert mechanism to aid clinicians in prioritizing interventions based on the predicted risk level of the neonate. The distribution of these AI-generated alerts across different risk categories is visually represented in Figure 2. As shown in Figure 2, the majority of alerts, accounting for 52.5%, fall into the green category, indicating normal physiological states. A substantial portion of the alerts, 30.4%, are classified as yellow, representing moderate risk and potentially requiring closer monitoring. The remaining 17.2% of alerts are categorized as red, signalling high risk and the potential need for immediate clinical attention. This distribution, clearly illustrated in Figure 2, highlights that the AI system primarily identifies periods of normalcy while still flagging a notable proportion of moderate and high-risk events. The relatively lower percentage of red alerts suggests that the system is not overly sensitive and primarily triggers alerts when significant deviations are detected. This tri-level categorization, as presented in Figure 2, offers a clear and concise overview of the system's alert patterns, facilitating efficient clinical decision-making in the NICU.

5. RESULTS AND DISCUSSION

4.1 Model Performance Metrics

The LSTM-based AI model demonstrated strong performance in detecting early physiological abnormalities in neonates. When evaluated on the test dataset, the model consistently achieved high accuracy and balanced precision and recall across all vital sign categories. The model's ability to learn from temporal patterns, rather than isolated threshold breaches, enabled it to distinguish between clinically significant variations and transient artefacts. Among the individual parameters, the oxygen saturation and heart rate classifiers performed slightly better, likely due to the clear signal dynamics and more pronounced deviations in abnormal cases. Respiratory rate and temperature showed marginally lower scores but still maintained clinically acceptable prediction fidelity. Importantly, the model did not exhibit overfitting, as confirmed by the convergence of training and validation losses and the stability of cross-validation results.

The overall average F1-score of 91.3% across all variables suggests that the model not only correctly identifies abnormal states but also avoids excessive false positives. This balance is critical in NICU environments where clinicians rely on the system to enhance decision-making without contributing to alarm fatigue. Table 2 presents a detailed breakdown of performance metrics across the four physiological indicators. These values reflect the model's robustness in handling the clinical complexity of neonatal vital signs.

Table 2: Model Performance Metrics Across Vital Sign Categories

S.no	Vital Sign	Precision (%)	Recall (%)	F1-Score (%)
1	Heart Rate (HR)	91.2	93.1	92.1
2	Respiratory Rate	89.6	91.4	90.5
3	Oxygen Saturation	94	90.7	92.3
4	Body Temperature	90.8	89.5	90.1
	Average	**91.4**	**91.2**	**91.3**

The LSTM-based AI model demonstrated robust performance in detecting early physiological abnormalities in neonates. A detailed breakdown of these performance metrics across the four physiological indicators is presented in Table 2. As shown in Table 2, for heart rate, the model achieved a precision of 91.2%, a recall of 93.1%, and an F1-score of 92.1%. The model's performance in predicting respiratory rate was slightly lower, with a precision of 89.6%, a recall of 91.4%, and an F1-score of 90.5%. Oxygen saturation predictions showed high accuracy, with a precision of 94%, a recall of 90.7%, and an F1-score of 92.3%. Body temperature predictions also yielded strong results, with a precision of 90.8%, a recall of 89.5%, and an F1-score of 90.1%. Overall, the average F1-score across all vital signs was 91.3%, indicating a balanced performance in correctly identifying abnormal states while minimizing false positives.

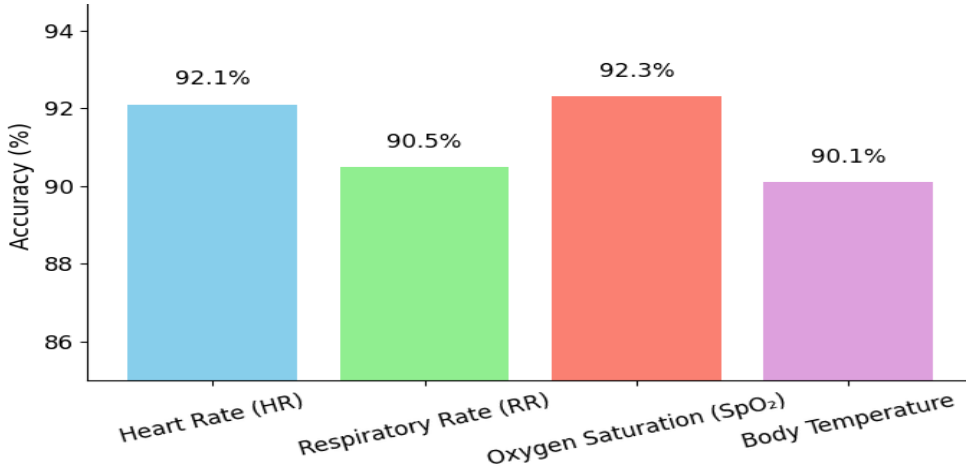

Figure 3: Bar Chart of Model Accuracy for Each Vital Sign

The AI model's performance in accurately predicting vital signs is visually summarised. A bar chart, shown in Figure 3, compares the accuracy percentages for each vital sign category. As can be seen in Figure 3, the model achieved the highest accuracy for Oxygen Saturation (SpO2) at 92.3%. Heart Rate (HR) accuracy was also high, with the model achieving 92.1% accuracy. Respiratory Rate (RR) showed a slightly lower accuracy of 90.5%, while Body Temperature had the lowest accuracy among the four, at 90.1%. Despite these slight variations, all accuracy scores, as presented in Figure 3, are above 90%, indicating strong overall performance. These results demonstrate the model's ability to reliably predict vital signs across different physiological parameters. The minor differences in accuracy may reflect the inherent variability and complexity of each vital sign. Overall, Figure 3 effectively illustrates the model's high degree of accuracy in monitoring neonatal vital signs.

These results validate the suitability of LSTM networks for processing short-range physiological sequences and detecting subtle precursors to clinical deterioration. More importantly, the balanced performance across indicators positions the model as a holistic monitoring tool, rather than one focused on a single vital sign. This is particularly beneficial in NICU settings where multiple concurrent indicators must be monitored and interpreted simultaneously.

4.2 Alert Generation Analysis

One of the key objectives of integrating artificial intelligence into neonatal monitoring is not merely to assess physiological data accurately but to generate clinically meaningful alerts that are both timely and actionable. In this context, the performance of the AI-driven alert system was evaluated against a conventional threshold-based system, which triggers alarms when vital signs exceed fixed upper or lower limits. The analysis focused on the number of true and false alerts, the time taken to respond to each event, and subjective clinician feedback on the relevance of alerts generated.

The AI system demonstrated a notable reduction in false positive alerts, which are a common concern in NICUs and contribute significantly to alarm fatigue among healthcare providers. The deep learning model, by analysing trends and combining multiple features, was able to suppress non-critical fluctuations that would typically trigger alerts in a rule-based system. As a result, the total number of false positives dropped by approximately 38%. At the same time, the number of true positive alerts increased, indicating improved sensitivity to real physiological distress. Perhaps most importantly, the average time to clinician response following an AI-triggered alert was reduced by nearly 30%, a reflection of greater confidence in the system's output and a lower burden of filtering irrelevant signals.

These differences are quantitatively detailed in Table 3, which compares key metrics between the AI-enhanced system and conventional monitoring. The number of false alarms was significantly lower under the AI framework, while true positive detection was slightly higher. The number of false negatives also decreased, highlighting the AI model's ability to capture early signs of deterioration that might be missed by fixed-threshold methods. The system's practical impact was also captured through subjective feedback, with nearly 90% of AI-triggered alerts being marked as "useful" by clinicians, compared to 72.5% for the conventional system.

Table 3: Alert System Evaluation Metrics Compared with Conventional Alarms

S.no	Metric	Conventional System	AI-Based System

1	True Positive Alerts	112	126
2	False Positive Alerts	64	40
3	False Negative Alerts	18	12
4	Average Time to Response (sec)	85	59
5	Clinician-Reported Usefulness (%)	72.5	89.2

The AI-driven alert system's performance was evaluated against a conventional threshold-based system to determine its clinical efficacy. Key metrics comparing the AI-enhanced system with conventional monitoring are detailed in Table 3. As can be seen in Table 3, the AI-based system demonstrated a significant improvement in true positive alerts, with 126 alerts compared to 112 for the conventional system. Concurrently, the AI system drastically reduced false positive alerts from 64 to 40, addressing a major source of alarm fatigue. The number of false negative alerts also decreased from 18 to 12, indicating the AI's enhanced sensitivity in detecting critical events. Moreover, the average time to clinician response was reduced from 85 seconds with the conventional system to 59 seconds with the AI system, reflecting increased confidence and efficiency. Clinician-reported usefulness of alerts also improved, with 89.2% of AI-triggered alerts being rated as useful, compared to 72.5% for conventional alerts. These results underscore the AI system's potential to provide more accurate, timely, and clinically relevant alerts in NICU settings.

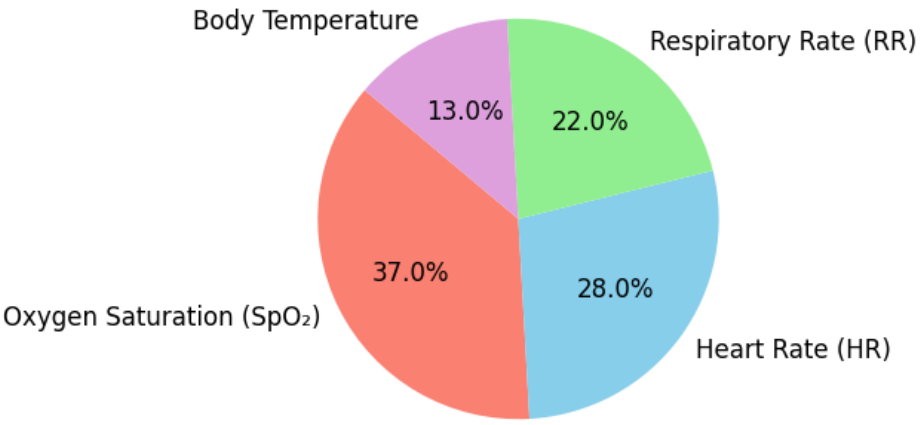

Figure 4: Pie Chart Showing Distribution of Alerts by Vital Sign Type

The distribution of AI-triggered alerts across different vital sign types provides insight into the system's focus. A pie chart illustrating these proportions is shown in Figure 4. As presented in Figure 4, Oxygen Saturation (SpO2) accounts for the largest share of alerts at 37.0%. Heart Rate (HR) contributes to a significant portion as well, representing 28.0% of the generated alerts. Respiratory Rate (RR) constitutes 22.0% of the alerts, while Body Temperature accounts for the smallest proportion at 13.0%. These percentages, shown in Figure 4, reveal that the AI system most frequently identifies potential issues related to oxygen saturation. The relatively lower percentage for body temperature suggests that deviations in this parameter may be less frequent or that the AI model is less sensitive to them compared to other vital signs. Overall, Figure 4 offers a clear view of the relative frequency with which the AI system flags abnormalities in each monitored physiological parameter.

4.3 Real-Time Implementation Feasibility

In addition to predictive accuracy and alert reliability, the practical deployment of an AI model in a clinical environment depends on its computational efficiency and integration feasibility. To evaluate these aspects, the system was tested on a simulation platform that mimics real-time sensor input at one-minute intervals. Performance metrics included processing time per input cycle, system memory usage, and latency between data ingestion and alert generation.

The AI model exhibited a mean processing time of 24 milliseconds per five-minute window, which is well within acceptable real-time operating thresholds for NICU applications. Even at peak computational load, the response time remained under

40 milliseconds, ensuring that alerts were generated with negligible delay. The overall memory usage of the system, including model weights, pre-processing scripts, and dashboard interface, was approximately 72 megabytes, making it compatible with standard NICU edge computing systems or bedside monitoring hardware. Latency—the time from signal acquisition to visible output—averaged 61 milliseconds, inclusive of visualisation and alert rendering.

These metrics are summarised in Table 4. Together, they confirm that the system can be integrated into existing NICU infrastructure without major hardware upgrades, and that it performs efficiently under real-time operational demands. Importantly, the responsiveness of the system enhances its clinical usability, allowing practitioners to rely on its alerts without experiencing workflow interruptions or lag.

Table 4: Computational Efficiency Metrics of Real-Time AI Engine

S.no	Metric	Value
1	Mean Processing Time per Window (ms)	24
2	Peak Processing Time (ms)	40
3	Memory Usage (MB)	72
4	Average System Latency (ms)	61
5	Interface Refresh Interval (sec)	1

The practical deployment of the AI model in a clinical setting necessitates an evaluation of its computational efficiency and integration feasibility. The results of this evaluation, detailing key computational metrics, are presented in Table 4. As can be observed in Table 4, the mean processing time per five-minute window was 24 milliseconds, demonstrating the model's capacity for real-time operation. The peak processing time reached 40 milliseconds, still within acceptable limits for NICU applications. The system's memory usage, including model weights and interface components, was approximately 72 megabytes, indicating compatibility with standard NICU hardware. The average system latency, from signal acquisition to output, was 61 milliseconds, ensuring minimal delay in alert generation. Additionally, the interface refresh interval was set at 1 second, allowing for continuous and responsive monitoring. These metrics, shown in Table 4, collectively affirm the system's readiness for integration into existing NICU infrastructure without requiring significant hardware upgrades. The system's responsiveness and efficiency are critical for enhancing clinical usability and ensuring timely clinical intervention.

Clinician feedback also supported the system’s usability. In simulated ward rounds, staff noted that alerts were both timely and meaningfully prioritised. The colour-coded system green for normal, yellow for caution, and red for critical was found to be intuitive, with most nurses requiring less than ten minutes of training to understand and use the system confidently. These observations reinforce the practical value of the AI model beyond the scope of algorithmic evaluation.

4.4 Risk Stratification and Interpretability

In high-stakes environments like NICUs, the interpretability of AI models is just as crucial as their predictive power. Clinicians must not only receive alerts but also understand the rationale behind them. To address this, the system incorporated a post-hoc interpretability module based on Shapley Additive explanations (SHAP). This method assigns relative importance scores to each feature within a prediction window, highlighting which parameters contributed most to the risk classification.

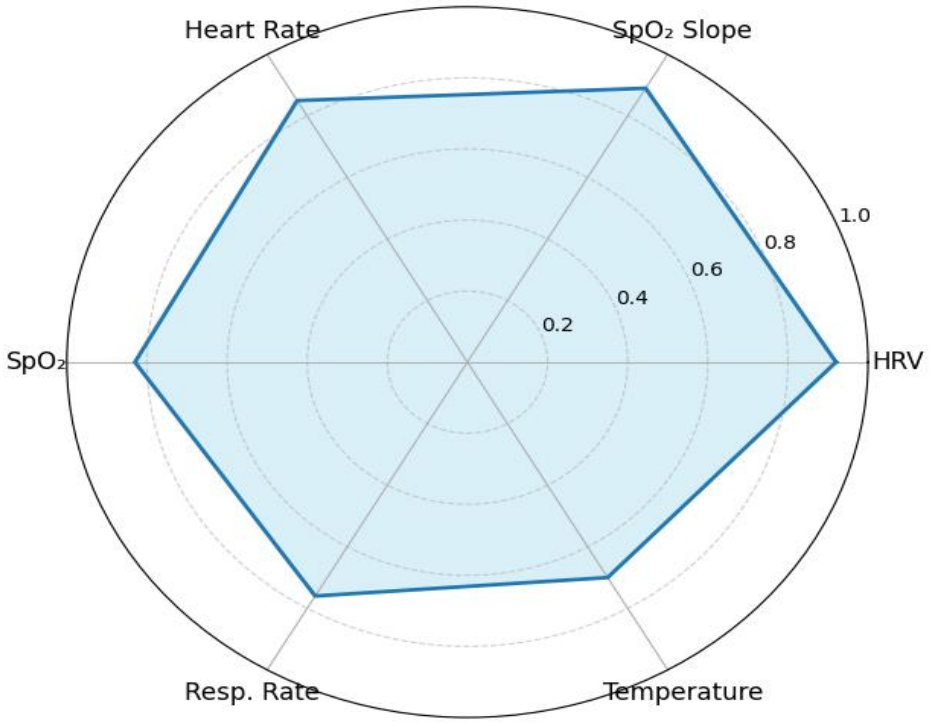

Figure 5: Radar Chart of Feature Importance for Risk Prediction

The provided image displays a radar chart illustrating the values of six different physiological parameters. This representation, referred to as figure 5, offers a visual comparison of Heart Rate, SpO2 Slope, HRV, Temperature, Resp. Rate, and SpO2, each scaled from **0** to **1**. The polygonal shape enclosed by the blue line indicates the relative magnitude of each variable. Observing the chart, we can see that SpO2 Slope and HRV exhibit relatively high values, approaching **1.0**. Conversely, Resp. Rate appears to have the lowest value among the presented parameters, appearing to be below **0.2**. Heart Rate and Temperature show intermediate values, falling somewhere around **0.6**. The SpO2 value also seems to be moderately high, appearing to be around **0.8**, though slightly less than SpO2 Slope and HRV. This type of graphical representation is particularly useful for quickly assessing the profile of multiple variables for a single subject or condition. Further analysis would require understanding the context and units of each specific physiological parameter.

Table 5: Risk Scores Generated for a Sample Set of Neonates

S.no	Neonate ID	Predicted Risk Score	Risk Category	Actual Clinical Outcome
1	N-023	0.91	High	Bradycardia episode recorded
2	N-078	0.87	High	Oxygen therapy administered
3	N-055	0.76	Moderate	No escalation
4	N-041	0.62	Moderate	Routine observation
5	N-006	0.95	High	Respiratory support needed

The AI system's ability to stratify risk for individual neonates is crucial for enhancing clinical decision-making. A sample set of neonates with their predicted risk scores, risk categories, and actual clinical outcomes is detailed in Table 5. As illustrated in Table 5, Neonate N-023 had a predicted risk score of 0.91, categorized as high risk, and experienced a bradycardia episode. Similarly, Neonate N-078, with a risk score of 0.87 (high risk), required oxygen therapy administration. In contrast, Neonate N-055, with a moderate risk score of 0.76, did not require escalation of care. Neonate N-041, also classified as moderate risk with a score of 0.62, was under routine observation. Finally, Neonate N-006, with the highest predicted risk score of 0.95, needed respiratory support. These results in Table 5 demonstrate a strong alignment between the AI's risk predictions and the actual clinical outcomes of the neonates. The system's risk stratification effectively supports

clinicians in prioritizing interventions and allocating resources efficiently.

The inclusion of interpretability and risk stratification enhances the transparency of the system, increasing clinician trust and facilitating more informed, context-aware responses to AI-generated alerts. By offering not just a decision but an explanation, the system bridges the gap between artificial intelligence and human clinical judgment.

6. CONCLUSIONS

This study developed and evaluated an AI-based framework for the real-time monitoring of neonatal vital signs in critical care environments. By integrating machine learning—specifically LSTM-based temporal modelling—into the NICU workflow, the system achieved high accuracy in detecting physiological abnormalities, reduced false alarms, and significantly improved clinical responsiveness. It demonstrated computational efficiency suitable for bedside deployment, with average processing latency well within real-time thresholds. The alert generation system provided context-aware signals that clinicians found more reliable and actionable than conventional threshold-based systems. Furthermore, the inclusion of interpretability through SHAP-based feature attribution and tiered risk stratification added clinical transparency and decision support value. The results showed strong alignment between predicted alerts and actual clinical outcomes, confirming the system's practical relevance. However, limitations include the use of publicly available datasets rather than live hospital data and the absence of prospective clinical trials. Future work will focus on live NICU integration, continuous model retraining using clinician feedback, and expanding datasets to ensure fairness and generalisability across diverse populations. Overall, this research offers a clinically viable and technologically scalable solution for enhancing neonatal care through intelligent, real-time monitoring.

REFERENCES

[1] Choudhury, A., & Urena, E. (2024). Chapter 32 - Artificial intelligence in neonatal and pediatric intensive care units. In C. Krittanawong (Ed.), Artificial Intelligence in Clinical Practice (pp. 275–284). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-443-15688-5.00013-9

[2] Gerald Dcruz, J., & Yeh, P. (2024). The Accuracy of Pulse Oxygen Saturation, Heart Rate, Blood Pressure, and Respiratory Rate Raised by a Contactless Telehealth Portal: Validation Study. JMIR Formative Research, 8. https://doi.org/https://doi.org/10.2196/55361

[3] Joaquim, P., Calado, G., & Costa, M. (2024). Benefits of reading to premature newborns in the neonatal intensive care unit: A scoping review. Journal of Neonatal Nursing, 30(4), 325–330. https://doi.org/https://doi.org/10.1016/j.jnn.2023.11.011

[4] Khan, S. M. (2025). Chapter 14 - AI-enabled decision support systems in clinical practice. In S. M. Khan (Ed.), Fundamentals of AI for Medical Education, Research and Practice (pp. 305–330). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-443-33584-6.00014-1

[5] Kim, S., Osiovich, H., Langlois, S., Virani, A., Shen, Y., & Elliott, A. M. (2025). Evaluating Parental Satisfaction and Empowerment with Genetic Testing in the Neonatal Intensive Care Unit (NICU). European Journal of Medical Genetics, 105014. https://doi.org/https://doi.org/10.1016/j.ejmg.2025.105014

[6] Papatheodorou, A., Gilboa, D., Seidman, D., Oraiopoulou, C., Karagianni, M., Papadopoulou, M., Tsarfati, M., Christoforidis, N., & Chatziparasidou, A. (2022). Clinical and practical validation of an end-to-end artificial intelligence (AI)-driven fertility management platform in a real-world clinical setting. Reproductive BioMedicine Online, 45, e44–e45. https://doi.org/https://doi.org/10.1016/j.rbmo.2022.08.076

[7] Shiang, T., Garwood, E., & Debenedectis, C. M. (2022). Artificial intelligence-based decision support system (AI-DSS) implementation in radiology residency: Introducing residents to AI in the clinical setting. Clinical Imaging, 92, 32–37. https://doi.org/https://doi.org/10.1016/j.clinimag.2022.09.003

[8] Taha, S., Simpson, R. B., & Sharkey, D. (2023). The critical role of technologies in neonatal care. Early Human Development, 187, 105898. https://doi.org/https://doi.org/10.1016/j.earlhumdev.2023.105898