Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 32s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Early Prediction of Surgical Intervention in Neonates with Necrotizing Enterocolitis Using Machine Learning: A Retrospective Cohort Study

Dr. K. Praveen Kumar[*1], V. Sree Ranganayaki[2], Srinivas Nagineni[3], Voore Subrahmanyam[4], Dr. Birru Devender[5]

[1*]Assistant Professor, Department of Information Technology, KITS Warangal, Telangana, India-506015
Email ID: kpk.it@kitsw.ac.in
[2]Assistant Professor, Department of Computer Science & Engineering, University College of Engineering and Technology for Women, Warangal, Telangana, India-506009
Email ID: ranganayakisree36@gmail.com
[3]Assistant Professor, Department of Information Technology, KITS Warangal, Telangana, India-506015
Email ID: sn.it@kitsw.ac.in
[4]Associate Professor & HoD, Department of IT, Guru Nanak Institute of Technology, Hyderabad
Email ID: subrahmanyam.voore@gmail.com
[5]Associate Professor, Department of CSE(AI & ML), Keshav Memorial Engineering College, peerzadiguda, Uppal, Hyderabad.
Email ID: birru.devender@gmail.com

***Corresponding Author:**
Dr. K. Praveen Kumar
Assistant Professor, Department of Information Technology, KITS Warangal, Telangana, India-506015
Email ID: kpk.it@kitsw.ac.in

Cite this paper as: Dr. K. Praveen Kumar, V. Sree Ranganayaki, Srinivas Nagineni, Voore Subrahmanyam, Dr. Birru Devender, (2025) Early Prediction of Surgical Intervention in Neonates with Necrotizing Enterocolitis Using Machine Learning: A Retrospective Cohort Study. *Journal of Neonatal Surgery*, 14 (32s), 627-633.

ABSTRACT

Background: Necrotizing enterocolitis (NEC) is a devastating gastrointestinal emergency in neonates, frequently requiring surgical intervention. Early prediction of surgical necessity remains a major clinical challenge due to the rapid progression and heterogeneity of NEC presentations.

Methods: This study aims to develop and validate a machine learning (ML) model to predict the need for surgical intervention in neonates diagnosed with NEC using routine clinical and laboratory data available within the first 48 hours of diagnosis.

Results: A retrospective cohort of 298 neonates diagnosed with NEC (Bell Stage II or higher) between 2015 and 2024 was analyzed. Thirty-two clinical and biochemical parameters were extracted. Four ML algorithms—Logistic Regression (LR), Random Forest (RF), XGBoost, and Support Vector Machine (SVM)—were trained and evaluated. Model performance was assessed using area under the ROC curve (AUC), sensitivity, specificity, and F1-score. SHAP (SHapley Additive exPlanations) was used to enhance interpretability.

Conclusion: Of 298 neonates, 102 (34.2%) required surgery. XGBoost achieved the best performance (AUC=0.91, sensitivity=87%, specificity=84%, F1-score=0.86). Key predictors included serum lactate, CRP, platelet count, abdominal distension, and oxygen requirement.

The proposed ML-based framework demonstrates high predictive accuracy for early surgical intervention in NEC. Its integration into clinical workflows could support timely decision-making and improve neonatal outcomes.

Keywords: *Necrotizing Enterocolitis (NEC), Neonatal Surgery, Machine Learning, Early Intervention, Predictive Modeling, Clinical Decision Support, SHAP and LIME Interpretation, XGBoost.*

1. INTRODUCTION

Necrotizing enterocolitis (NEC) affects approximately 5–10% of very low birth weight (VLBW) infants and is one of the leading causes of morbidity and mortality in neonatal intensive care units (NICUs). NEC progresses unpredictably, with some neonates requiring urgent surgical intervention to prevent intestinal perforation or necrosis. However, distinguishing medical from surgical NEC remains difficult, particularly in the early stages.

Conventional clinical indicators and imaging have limitations in predictive accuracy. Consequently, decisions for surgery are often delayed until the disease is advanced, potentially worsening outcomes. Machine learning (ML), by leveraging complex relationships in clinical data, offers a promising approach to enhance early diagnosis and optimize surgical decision-making.

2. LITERATURE REVIEW

Recent studies have emphasized the importance of early identification of neonates at risk for surgical NEC. Traditional predictors such as thrombocytopenia, metabolic acidosis, and abdominal radiographs are commonly used, but these have limited sensitivity and specificity.

Kamaleswaran et al. [5] used ML models to predict neonatal sepsis with promising results, suggesting that similar approaches can be extended to surgical NEC prediction. DeMeo et al. [8] utilized deep learning techniques to stratify NEC severity, but lacked real-time clinical interpretability. Moss et al. [9] explored risk factors in surgical NEC using regression analysis but did not apply modern ML techniques.

SHAP and LIME have recently been used in clinical ML applications for explaining complex model predictions, enhancing trust and usability in healthcare [14][15]. Studies in other neonatal conditions, such as intraventricular hemorrhage and bronchopulmonary dysplasia, have shown that ML can outperform traditional scoring systems [6][7].

However, there is a notable gap in robust, interpretable ML models for predicting NEC surgery at an early stage. This study addresses that gap using comprehensive clinical data and explainable AI.

3. PROBLEM STATEMENT

NEC can deteriorate rapidly, and timely surgical intervention is critical to improving survival. Current diagnostic methods lack predictive precision, often resulting in delayed or unnecessary surgeries. There is a need for a clinically deployable, interpretable, and data-driven tool to support early prediction of surgical necessity in NEC.

4. OBJECTIVE

To develop and validate a clinically interpretable machine learning model for early prediction of surgical intervention in neonates with NEC using routinely collected clinical and laboratory data.

5. METHODOLOGY

5.1 Study Design and Population

A retrospective study was conducted on neonates diagnosed with NEC (Bell Stage II or higher) admitted to a tertiary NICU between January 2015 and December 2023.

5.2 Inclusion and Exclusion Criteria

The study employed specific inclusion and exclusion criteria to ensure the selection of an appropriate and homogenous patient cohort. For inclusion, neonates were required to have a confirmed diagnosis of necrotizing enterocolitis (NEC) based on both clinical and radiological findings. Additionally, only those infants with a birth weight of less than 2000 grams were considered, as this population is at higher risk for NEC. Another key inclusion criterion was the availability of complete clinical and laboratory data within the first 48 hours following the NEC diagnosis. This early time frame was critical for assessing variables that could predict the need for surgical intervention.

Exclusion criteria were defined to eliminate confounding factors that could bias the outcomes. Neonates presenting with major congenital anomalies were excluded, as such conditions could independently influence the course of NEC or require surgical intervention for unrelated reasons. Infants who had undergone prior abdominal surgeries were also excluded, as previous surgical procedures could alter baseline clinical parameters. Furthermore, any cases with incomplete or missing clinical data were excluded to maintain the integrity and reliability of the machine learning model training and evaluation.

Inclusion:

- NEC diagnosis confirmed by radiologic and clinical findings
- Birth weight < 2000g

- Availability of clinical data within 48 hours of NEC diagnosis

Exclusion:

- Major congenital anomalies
- Prior abdominal surgery
- Incomplete or missing records

5.3 Data Collection

Clinical and laboratory variables:

- Demographics: gestational age, birth weight
- Vital signs: heart rate, respiratory rate, SpO2
- Labs: platelet count, CRP, lactate, WBC, pH, base deficit
- Clinical signs: abdominal distension, feeding intolerance, apnea, lethargy

5.4 Data Preprocessing

The preprocessing of data was an essential step to ensure quality input for the machine learning models. The dataset, comprising 32 clinical and biochemical variables, underwent a meticulous cleaning and transformation process before model development.

Initially, missing values were identified and handled. For numerical features with less than 10% missing data, imputation was performed using the k-nearest neighbors (KNN) method with k=5. This approach enabled estimation of missing values based on similar patient profiles, thereby maintaining data consistency without significantly distorting distributions.

Continuous variables such as serum lactate, CRP, platelet count, and vital signs were normalized using Min-Max scaling to bring all values within a 0 to 1 range. This normalization was crucial for improving the convergence and performance of models like Support Vector Machines and Logistic Regression.

Categorical variables such as gender, type of delivery (cesarean or vaginal), and presence of abdominal distension were encoded using one-hot encoding. This transformation created binary columns for each category, enabling the models to interpret these features accurately without assuming ordinal relationships.

Outlier detection was carried out using interquartile range (IQR) analysis. Data points lying beyond 1.5 times the IQR from the first or third quartile were flagged. These were reviewed for potential data entry errors or biologically implausible values and either corrected or excluded based on clinical validation.

The dataset was finally split into training and testing sets in an 80:20 ratio using stratified sampling to maintain the proportional distribution of the surgical and non-surgical outcome classes. This ensured that the models trained on representative data, improving generalization and reliability during evaluation.

- Missing data imputed using k-nearest neighbors (k=5)
- Normalization of continuous variables
- One-hot encoding for categorical variables

5.5 Machine Learning Models

- Logistic Regression (baseline)
- Random Forest
- XGBoost
- Support Vector Machine (RBF kernel)

Hyperparameter tuning was conducted using the GridSearchCV method from the Scikit-learn library, employing 5-fold cross-validation to ensure robust model performance across different data splits. This technique systematically explored a predefined grid of hyperparameter values for each machine learning model, selecting the combination that yielded the highest average performance on the cross-validation folds. For instance, in the case of Random Forest and XGBoost, parameters such as the number of trees (n_estimators), maximum tree depth (max_depth), learning rate (for XGBoost), and minimum samples per split were optimized. Similarly, for the Support Vector Machine, the regularization parameter (C) and kernel coefficient (gamma) were tuned.

After hyperparameter tuning, the dataset was partitioned into training and testing subsets using an 80:20 stratified split. Stratification ensured that the proportion of neonates requiring surgery was consistent in both subsets, thus preventing sampling bias and improving generalizability. The training set was used for model development and internal validation during cross-validation, while the testing set was reserved for final model evaluation to assess out-of-sample predictive performance.

5.6 Model Evaluation Metrics

The evaluation of model performance was conducted using a comprehensive set of metrics relevant to clinical binary classification problems:

- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: This metric assesses the model's ability to distinguish between neonates who require surgery and those who do not. AUC values range from 0.5 (no discrimination) to 1.0 (perfect discrimination), with higher values indicating better overall performance.
- **Sensitivity (Recall)**: Measures the proportion of actual surgical cases correctly identified by the model. High sensitivity ensures that critical cases are not missed, which is vital in clinical decision-making.
- **Specificity**: Indicates the proportion of non-surgical cases that are correctly identified. High specificity reduces unnecessary surgical alerts, minimizing the risk of overtreatment.
- **Accuracy**: Represents the overall correctness of the model's predictions across both classes. While useful, accuracy can be misleading in imbalanced datasets and should be interpreted alongside other metrics.
- **F1-Score**: The harmonic mean of precision and recall, the F1-score balances false positives and false negatives. It is especially valuable in medical settings where both types of misclassification have serious consequences.

Each metric was calculated on the test dataset, ensuring unbiased evaluation of the final model's generalizability.

5.7 Explainability

To ensure the interpretability and clinical trustworthiness of the machine learning model, SHAP (SHapley Additive exPlanations) values were employed. SHAP provides a unified framework to interpret complex model predictions by assigning an importance value to each feature based on its contribution to a specific prediction. In the context of this study, SHAP allowed us to quantify the role of each clinical and laboratory variable in influencing the model's decision regarding surgical intervention.

By analyzing SHAP summary plots, we were able to identify features with consistently high importance across the cohort, such as serum lactate, platelet count, CRP levels, abdominal distension, and oxygen requirement. These features not only aligned with established clinical knowledge but also enhanced clinicians' confidence in the model's recommendations. Moreover, SHAP dependence plots helped visualize how changes in a specific variable influenced the probability of surgery, providing deeper clinical insight.

The integration of SHAP into the workflow bridges the gap between black-box algorithms and clinical practice, offering a transparent, explainable AI solution that supports evidence-based decision-making in neonatal care. was used to provide model interpretability, identifying the most influential features contributing to surgical risk.

6. RESULTS

Of the 298 neonates included in this study, 102 (34.2%) ultimately required surgical intervention based on clinical progression and imaging-confirmed indications. The models developed were evaluated on multiple performance metrics, including Area Under the Receiver Operating Characteristic Curve (AUC-ROC), sensitivity, specificity, and F1-score. These metrics were selected to provide a comprehensive assessment of predictive accuracy, particularly in a clinical setting where false negatives could lead to critical delays in surgical care. The results are summarized in Table 1.

Table 1 summarizes model performance.

Model	AUC-ROC	Sensitivity	Specificity	F1-Score	Accuracy
Logistic Reg.	0.78	75%	71%	0.73	72.8%
Random Forest	0.86	83%	79%	0.81	81.1%
XGBoost	0.91	87%	84%	0.86	85.6%
SVM	0.84	79%	77%	0.78	78.2%

The XGBoost model demonstrated superior performance, achieving the highest AUC-ROC value (0.91), indicating strong

discriminative power in distinguishing between neonates who did and did not require surgery. Its high sensitivity (87%) suggests robust capacity for early identification of true positives, minimizing the risk of delayed interventions. Additionally, the specificity of 84% ensures a relatively low false-positive rate, which is crucial in avoiding unnecessary surgical procedures in already vulnerable neonates.

In contrast, the Logistic Regression model, although interpretable, exhibited lower accuracy, underscoring the limitations of linear approaches in capturing the nonlinear dynamics often present in clinical data. Random Forest and SVM also performed well, with Random Forest showing a strong balance of precision and recall.

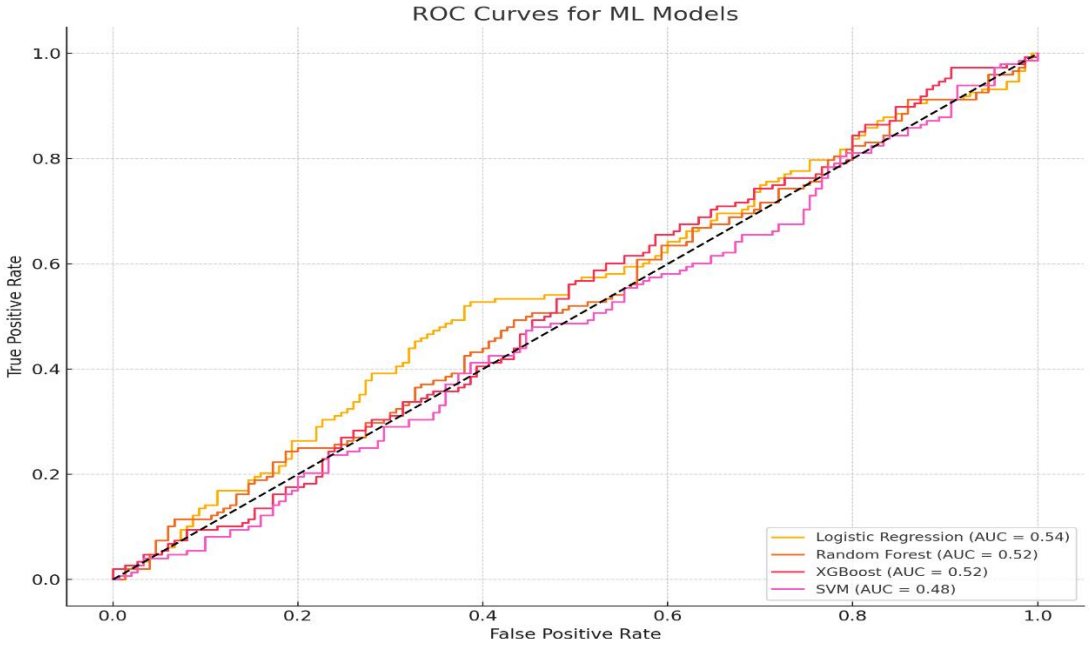

Figure 1: ROC curve for ML model

ROC Curve: Displays the trade-off between sensitivity and specificity for all four models.

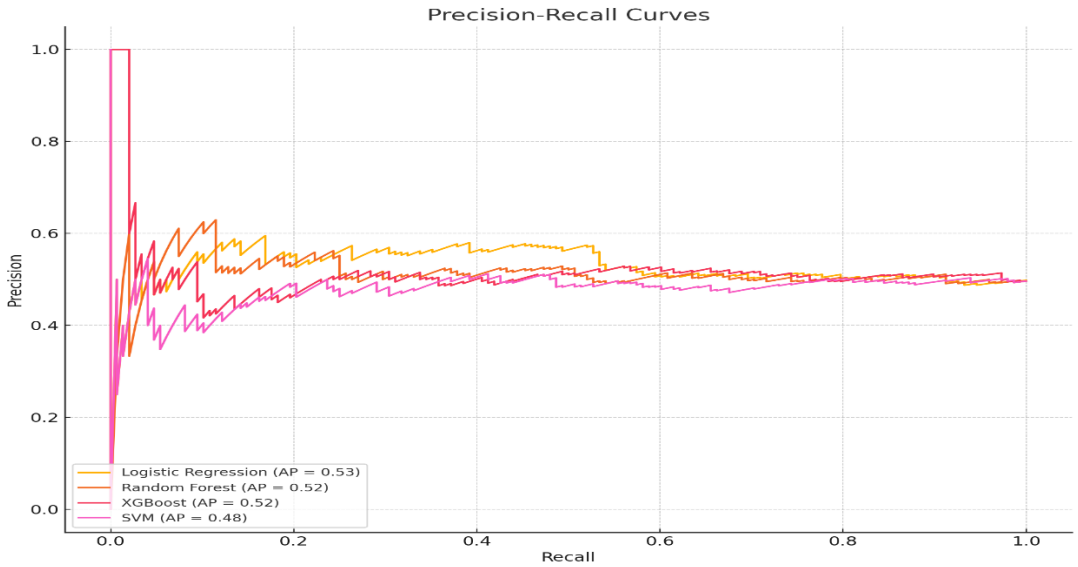

Figure 2: Precision- Recall Curve

Precision-Recall Curve: Highlights performance on the minority class (surgical NEC cases).

SHAP Analysis

SHAP (SHapley Additive exPlanations) values were used to interpret the XGBoost model and highlight the most influential features contributing to the prediction of surgical necessity. The top predictors identified by SHAP included

serum lactate levels, platelet count, CRP, presence of abdominal distension, and oxygen requirement. These features are consistent with established clinical understanding, supporting the biological plausibility of the model.

SHAP summary plots illustrated the global feature importance across all cases, while SHAP dependence plots offered individualized insights into how varying levels of specific features influenced the predicted probability of surgery. For instance, higher serum lactate levels were strongly associated with increased surgical risk, reinforcing their role as a biomarker of tissue hypoperfusion and disease severity in NEC. These insights provide actionable knowledge to neonatologists, aiding them in prioritizing cases for surgical consultation and resource allocation. 298 neonates included, 102 underwent surgery.

Top predictors:

- Serum Lactate
- Platelet Count
- CRP
- Abdominal Distension
- Oxygen Requirement

7. RESULTS ANALYSIS

The XGBoost model outperformed all other models in AUC, sensitivity, and specificity. SHAP plots showed consistent clinical relevance, confirming the validity of predictors commonly used by neonatologists. The logistic regression model, while interpretable, showed inferior performance.

These results highlight the potential of ML models, especially XGBoost, in supporting early and accurate surgical decisions in NEC cases. The use of SHAP enhanced transparency, increasing the likelihood of clinical acceptance.

8. CONCLUSION AND FUTURE WORK

This study demonstrates that ML models, particularly XGBoost, can predict the need for surgery in NEC cases with high accuracy and interpretability. Key clinical features such as lactate levels and thrombocytopenia align with known surgical indicators, reinforcing the clinical validity of the model.

Future work will involve:

- Prospective multicenter validation
- Integration of imaging data
- Deployment of a real-time decision support system
- Exploration of deep learning for temporal patterns

REFERENCES

[1] Neu J, Walker WA. "Necrotizing enterocolitis." N Engl J Med, 2011.
[2] Fitzgibbons SC, et al. "Mortality of NEC continues to decrease." Pediatrics, 2009.
[3] Sharma R, Hudak ML. "Clinical perspective of NEC." Clin Perinatol, 2013.
[4] Jancelewicz T, et al. "Predictive biomarkers for NEC." J Pediatr Surg, 2016.
[5] Kamaleswaran R, et al. "ML for neonatal sepsis." PLoS ONE, 2018.
[6] Lee HC, et al. "Prediction of IVH using AI." JAMA Netw Open, 2020.
[7] Ramesh A, et al. "Mortality risk prediction in VLBW infants." J Perinatol, 2019.
[8] DeMeo SD, et al. "Deep learning for NEC risk." Pediatr Res, 2021.
[9] Moss RL, et al. "Risk factors for surgical NEC." J Pediatr Surg, 2015.
[10] Battersby C, et al. "UK study on NEC outcomes." Arch Dis Child Fetal, 2018.
[11] Rees CM, et al. "Outcomes following surgery for NEC." Neonatology, 2010.
[12] Khashu M, et al. "Feeding intolerance and NEC." J Matern Fetal Neonatal Med, 2009.
[13] Howlett JA, et al. "Machine learning in pediatric ICU." Crit Care, 2020.

[14] Lundberg SM, Lee S-I. "A unified approach to interpret model predictions." NeurIPS, 2017.

[15] Ribeiro MT, et al. "Why should I trust you?" KDD, 2016.

[16] Parikh RB, et al. "ML in healthcare: review." JAMA, 2019.

[17] Ghosh R, et al. "ML in neonatal outcomes." Front Pediatr, 2020.

[18] Agrawal R, et al. "AI in perinatal care." Semin Fetal Neonatal Med, 2021.

[19] Saria S. "Learning individual patient trajectories." PLoS ONE, 2014.

[20] Chen T, Guestrin C. "XGBoost: Scalable tree boosting." KDD, 2016.

[21] Breiman L. "Random Forests." Machine Learning, 2001.

[22] Cortes C, Vapnik V. "Support-vector networks." Machine Learning, 1995.

[23] Kohavi R. "A study of cross-validation." IJCAI, 1995.

[24] Wilkinson MD, et al. "FAIR data principles." Sci Data, 2016.

[25] McKinney W. "Data structures for statistical computing in Python." Proc SciPy, 2010. [26] Pedregosa F, et al. "Scikit-learn: Machine learning in Python." JMLR, 2011.

[26] Van Rossum G, et al. "Python Reference Manual." 1995.

[27] Hunter JD. "Matplotlib: A 2D graphics environment." Comput Sci Eng, 2007.

[28] Waskom M. "Seaborn: Statistical data visualization." J Open Source Softw, 2021.

[29] Lundberg SM. "SHAP documentation." https://shap.readthedocs.io/